AD NUMBER

AD431994

NEW LIMITATION CHANGE

TO

Approved for public release, distribution unlimited

FROM

Distribution authorized to U.S. Gov't. agencies and their contractors; Administrative/Operational Use; NOV 1963. Other requests shall be referred to Office of Naval Research, Arlington, VA.

AUTHORITY

ONR ltr, 28 Jul 1977

SEL-63-099

Reproducing Distributions for Machine Learning

by
J. D. Spragins, Jr.

November 1963

Technical Report No. 6103-7

Prepared under
Office of Naval Research Contract
Nonr-225(24), NR 373 360
Jointly supported by the U.S. Army Signal Corps, the
U.S. Air Force, and the U.S. Navy
(Office of Naval Research)

SYSTEMS THEORY LABORATORY

STANFORD ELECTRONICS LABORATORIES

STANFORD UNIVERSITY · STANFORD, CALIFORNIA

SEL-63-099

REPRODUCING DISTRIBUTIONS FOR MACHINE LEARNING

by

J. D. Spragins, Jr.

November 1963

Technical Report No. 6103-7

Prepared under
Office of Naval Research Contract
Nonr-225(24), NR 373 360
Jointly supported by the U.S. Army Signal Corps, the U.S. Air Force, and
the U.S. Navy (Office of Naval Research)

Systems Theory Laboratory
Stanford Electronics Laboratories
Stanford University Stanford, California

ABSTRACT

A model is proposed for learning the nature and value of an unknown parameter, or unknown parameters, in a probability distribution which forms part of a body of statistics related to some system or process. The model is Bayesian, involving the assumption of an a priori probability distribution over the possible values of the unknown parameters; the performance of experiments to gain information about the parameters; and the alteration of the a priori probabilities by Bayes' rule. In the limit, as the number of experiments approaches infinity, the a posteriori distribution in most cases encountered in practice approaches a delta function at the true values of the unknown parameters, so the system learns the values of the parameters exactly. The learning process developed in the paper is shown to be technically feasible if the a priori and a posteriori distributions are of the same form, with the learning accomplished by calculating new parameters for these distributions. It is shown that a necessary and sufficient condition for fulfillment of this feasibility criterion is for a sufficient statistic of fixed dimension to exist. If such a sufficient statistic exists, the a posteriori distributions may vary in form initially, but they eventually become of fixed form. The techniques developed indicate logical methods for choosing a priori probabilities and are applied in pattern recognition, estimation, and other problems.

CONTENTS

CONTENTS (Continued)

CONTENTS (Continued)

TABLES

ILLUSTRATIONS

LIST OF SYMBOLS

A	constant
a	amplification of radio channel
B	constant
C	constant
C_n	parameter in simple reproducing density for exponential distribution
$d(X)$	decision rule
$E[\cdot]$	statistical expectation
$E[Z]$	expectation of performance criterion Z
$E[Z\|\theta]$	conditional expectation of Z given θ, expressed as function of θ
$E[Z\|\Lambda_i, \ldots \Lambda_j]$	expectation of Z in light of observations $\Lambda_i, \ldots \Lambda_j$
$\hat{E}[p(\theta)\|\Lambda_i, \ldots \Lambda_j]$	expectation of $p(\theta)$ with respect to density $\hat{p}(\theta\|\Lambda_i, \ldots \Lambda_j)$
f	frequency
$f(\cdot)$	function
$g(\cdot)$	function
$h(\cdot)$	function
$I_0(x)$	modified zero order Bessel function of first kind
$I_1(x)$	modified first order Bessel function of first kind
K	constant
K_n	parameter in simple reproducing density for Rayleigh distribution
$\mathbf{K}$	covariance matrix for Gaussian distribution
$L(\omega,\hat{\omega})$	loss function for estimation problem
M_n	maximum absolute value of observation from rectangular distribution

LIST OF SYMBOLS (Continued)

$\mathbf{M}$	mean vector of Gaussian distribution
m	one-dimensional mean of Gaussian distribution
N_o	noise intensity
n	number of observations
P	parameter characterizing binomial distribution
$\hat{P}$	estimate of P
$P(\cdot)$	probability mass function
P_i	parameter characterizing multinomial distribution
P_{ii}	parameter characterizing binary Markov distribution
$P(i)$	probability of observing i^{th} pattern
$P(i\|X)$	conditional probability of i^{th} pattern given observation X
$P(\Lambda_i, \ldots \Lambda_j\|\theta)$	likelihood function with discrete random variable θ
$p(\cdot)$	probability density
$p(X)$	probability density of observation X
$p(X\|i)$	probability density of X given i^{th} pattern class
$p(\mathbf{X})$	probability density of vector-valued observation X
$p(\theta)$	a priori probability density assumed for θ
$p(\theta\|\Lambda_i, \ldots \Lambda_j)$	a posteriori density for θ in the light of the observations $\Lambda_i, \ldots \Lambda_j$
$\hat{p}(\theta\|\Lambda_i, \ldots \Lambda_j)$	"experimental portion" of $p(\theta\|\Lambda_i, \ldots \Lambda_j)$
$p(\Lambda_i, \ldots \Lambda_j\|\theta)$	likelihood function with continuous random variable θ
r	number of observations of given event with binomial, multinomial or binary Markov distribution

$r(\theta)$	non-negative integrable function of θ in composite reproducing density
R_n	parameter in composite reproducing density for complex Gaussian distribution
$T(\Lambda_i, \ldots \Lambda_j)$	sufficient statistic for θ as function of $\Lambda_i, \ldots \Lambda_j$
$t_i^{(n)}$	component of sufficient statistic for a posteriori observations
$t_i^{(-m,n)}$	component of sufficient statistic for combined a priori and a posteriori observations
v_n	sample scatter about mean for observations from Gaussian distribution
v_n^*	sample scatter about sample average for observations from Gaussian distribution
W	width of rectangular distribution
X	real-valued observation
$\mathbf{X}$	vector-valued observation
$\lvert\overline{X}_n\rvert$	parameter of simple reproducing density for complex Gaussian distribution
$\overline{\mathbf{X}}_n$	sample average of observations from Gaussian distribution
Z	random variable representing performance criterion
α	mean of Poisson distribution (also used as general parameter)
$\Gamma(x)$	gamma function
$\delta(x)$	Dirac delta function
δ_n	parameter in simple reproducing density for complex Gaussian distribution
ϵ	"is in" or "belongs to"
θ	random variable representing unknown parameter or parameters

LIST OF SYMBOLS (Continued)

Λ	set of observations
$\Lambda_1, \ldots \Lambda_n$	a posteriori set of observations
$\Lambda_{-m}, \ldots \Lambda_0$	a priori set of observations
λ	parameter characterizing exponential distribution
$\hat{\lambda}$	estimate of λ
$\boldsymbol{\mu}_n$	mean vector of Gaussian density assumed for learning mean of Gaussian distribution
ρ	inverse parameter for Rayleigh distribution
σ^2	variance of univariate Gaussian distribution (or corresponding parameter in other distributions)
τ	observation time for Poisson process
ϕ	phase shift in complex Gaussian distribution
$\Phi(x)$	Gaussian cumulative distribution function
$\boldsymbol{\phi}_n$	covariance matrix of Gaussian density assumed for learning mean of Gaussian distribution
ω	parameter to be estimated
$\hat{\omega}$	estimate of ω
$\omega = 2\pi f$	angular frequency

ACKNOWLEDGMENT

The work described in this report is a continuation of an investigation of the machine learning process begun by Professor D. J. Braverman of California Institute of Technology, Pasadena. The theory has been developed by several investigators at Stanford, especially Professor Norman Abramson and Dr. D. J. Keehn (now of International Telephone and Telegraph Federal Laboratories). Dr. Keehn's generalization of the learning theory furnished the immediate impetus leading to the work reported here.

In addition to the persons mentioned above, the author would like to thank Professor Thomas Kailath of Stanford for his valuable suggestions in connection with the writing of this report. The author would especially like to thank Professor Abramson for his guidance and help during the entire course of the investigations.

I. INTRODUCTION

A. PURPOSE

The purpose of the study described in this paper is to develop a model for a learning technique capable of utilizing and evaluating statistical information relating to a physical system or process. The model is to be applicable in situations where the form of the probability distributions describing a process is known, but where the values of some of the parameters involved in these distributions are unknown. The model is to be readily adaptable to construction of an actual learning machine or to simulation of such a machine on a digital computer.

It is expected that the results of the study will be useful in the design of complex multiple-element systems, including a variety of different types of communication systems.

B. BACKGROUND

Since the pioneering work of Shannon and Wiener in 1948-49 [Refs. 1-4], a large amount of research has been done on application of statistical techniques to design of communication systems. This research has been motivated by the realization that often only an approximate estimate of the conditions under which a communication system will be required to operate is available. Under these circumstances, designing the system so that its performance will be the best possible on the average appears more reasonable than attempting to optimize performance under specific conditions which may later turn out to be inapplicable.

To achieve the best possible average system performance, statistical techniques are applied. A specific criterion for judging system performance is defined; then the techniques of probability theory are utilized to see how well this criterion may be expected to be satisfied. Stating the matter in more mathematical language, excellence of system performance is judged by the statistical expectation of a random variable Z which represents the selected performance criterion. In some cases Z is a squared error term, in which case its statistical expectation

$E[Z]$ is the mean squared error; in other cases Z is the fraction of the time when a system makes an error, with $E[Z]$ the probability of error.

Although the mathematics involved are often complex, the application of the statistical criteria is in principle straightforward provided a body of statistics relating to the problem is available. The statistics can often be computed through a knowledge of the physical principles involved, or can be estimated accurately from experience. In some cases, however, the statistics are not accurately known and must be further investigated before any criteria or statistical expectations thereof can be established. This fact is responsible for much of the current emphasis on research in learning techniques.

In connection with a body of statistics, a learning technique may be defined as a procedure for evaluating experimental observations in order to gain information about parameters involved in the system or process to which the statistics apply. Throughout this report the term learning will be used in the restricted sense suggested by this definition, and only in this restricted sense. In view of the large amount of research currently being done on learning in biological systems, it should be pointed out that learning in the sense in which the term is used here may bear little resemblance to learning performed by biological systems.

C. METHOD OF APPROACH

In this investigation a possible model for the process of learning the values of unknown parameters in a body of statistics is developed. Although the proposed model is not the most general possible, it is general enough for most practical purposes. One important kind of a priori information is postulated: it is assumed that the forms of the probability distributions involved in the statistics are known, although some of the parameters of these distributions are unknown. This assumption is interpreted to mean that the physical process involved is known well enough to identify the type of probability density being dealt with, but not well enough to permit computation of

all the parameters for this density. This is a situation often occurring in practice; for example, it might be known that a probability density was multivariate Gaussian, but the mean vector or covariance matrix for this Gaussian density might not be known.

As a basic procedure it is assumed that the symbol θ represents some unknown parameter or parameters in one of the known probability densities. In order that the statistical expectation $E[Z]$ can be computed θ is treated as a random variable and an a priori probability density $p(\theta)$ is assumed over the range of its possible values.* The expectation $E[Z]$ is then determined from the standard statistical equation

$$E[Z] = \int E[Z|\theta]\, p(\theta)\, d\theta \qquad (1)$$

The learning model developed in this investigation is based on a series of modifications of Eq. (1). These modifications will be discussed in the next chapter.

*This so-called "Bayesian" technique of treating a fixed but unknown parameter as a random variable is common engineering practice, though frowned on by many statisticians. Even in statistical circles, however, the practice appears to be gaining wider acceptance [Refs. 5 and 6].

II. THE LEARNING MODEL

A. BASIC EQUATION

It has been shown that, for a body of statistics related to some physical process or system,

$$E[Z] = \int E[Z|\theta]\, p(\theta)\, d\theta \qquad (1)$$

where:

- θ = an unknown parameter or parameters in the probability distributions included in the statistics
- Z = a random variable representing a selected performance criterion
- $E[Z]$ = the statistical expectation of Z
- $p(\theta)$ = the a priori probability density function of θ [$p(\theta)$ or some information which may be utilized in choosing $p(\theta)$ is assumed to be known a priori*]
- $E[Z|\theta]$ = the conditional expectation of Z given θ (the expectation of Z is assumed to be known a priori as a function of θ; for any specific value of θ, $E[Z|\theta]$ is the value that would be used for $E[Z]$ if θ were known to have the postulated value).

In this investigation Eq. (1) is to be used as the basis for a learning model; however, modification of Eq. (1) is suggested by the fact that, if the value of θ were known more accurately, more confidence could be placed in the value of $E[Z]$.

B. LEARNING OBSERVATIONS

The obvious way to improve the extent of knowledge about θ is to perform an experiment, or a set of experiments, to gain information about the parameters. Let the set of outcomes of some such set of learning observations be designated by Λ_1. Λ_1 cannot be expected to tell

*One of the results of this investigation is to indicate ways of choosing $p(\theta)$ when this density is only approximately known.

exactly what the value of θ is, since it has been assumed that θ cannot be measured accurately; however, it is assumed that the probability density function of the learning observations is known as a function of θ. If the probability density function of the learning observations were not known, or if it were not a function of θ, there would be little to gain from performing the experiments. The probability density function of the learning observations is denoted by $p(\Lambda_1|\theta)$.*

In the present study it is also assumed that $E[Z|\theta]$ is independent of Λ_1. This may be interpreted as an assumption that Λ_1 is used only to improve the extent of knowledge about θ and does not influence the values of θ. (An example of an equivalent assumption is the assumption that inserting an ammeter in an electric circuit to measure the current does not change the magnitude of the current; any other assumption that the measurement of a quantity does not influence the magnitude of that quantity is also equivalent.)

C. MORE ACCURATE VERSION OF STATISTICAL EXPECTATION

The information is now available to compute a more accurate version of $E[Z]$. First, Bayes' rule** is applied to obtain

$$p(\theta|\Lambda_1) = \frac{p(\Lambda_1|\theta)\, p(\theta)}{\int p(\Lambda_1|\theta)\, p(\theta)\, d\theta} \qquad (2)$$

---

*A quantity of the form of $p(\Lambda_1|\theta)$, when treated as a function of θ, is often called a "likelihood" rather than a "probability density." As a function of Λ_1, for fixed θ, $p(\Lambda_1|\theta)$ has been defined to be a probability density. As a function of θ for fixed Λ_1, however, $p(\Lambda_1|\theta)$ is not a true probability density; although it satisfies one of the requirements for a probability density by being non-negative, it does not normally satisfy the requirement of integrating to one. In the subsequent discussion the term "probability density" will be used when quantities of the form of $p(\Lambda_1|\theta)$ are considered as functions of observations, while the term "likelihood" will be used when such quantities are considered as functions of θ.

**Bayes' rule is the standard equation for computing conditional probabilities. It may be found in any textbook on probability theory.

where

Λ_1 = the outcomes of a set of learning observations used to gain information about θ

$p(\Lambda_1|\theta)$ = probability density function of the learning observations Λ_1 (when treated as a function of Λ_1 for fixed θ)

= likelihood function of θ (when treated as a function of θ for fixed Λ_1)

(this quantity is assumed to be known as a function of both Λ_1 and θ; it is used as a likelihood function in Eq. (2))

$p(\theta)$ = a priori probability density function of θ

$p(\theta|\Lambda_1)$ = a posteriori probability density function of θ (this function is assumed to be evaluated in the light of Λ_1 by Eq. (2))

The new expectation for Z is then calculated as

$$E[Z|\Lambda_1] = \int E[Z|\theta]\, p(\theta|\Lambda_1)\, d\theta \qquad (3)$$

where:

$E[Z|\Lambda_1]$ = the statistical expectation of Z incorporating the information gained from the observations Λ_1

= the conditional expectation of Z given the observations Λ_1

$E[Z|\theta]$ = the conditional expectation of Z given θ, expressed as a function of θ, and assumed independent of Λ_1.

This calculation completes one stage of the learning process. A more accurate version of E[Z] has been obtained, but it may be desired to obtain a still more accurate version. This even more accurate version can be obtained by repeating the previous process. Another set, Λ_2, of learning observations is taken; $p(\theta|\Lambda_1,\Lambda_2)$ is computed by Bayes' rule; and this density is used to compute $E[Z|\Lambda_1,\Lambda_2]$. Then a third set, Λ_3, of learning observations is taken and the process is repeated. The progressively developing results of the learning process can be expressed in terms of the three sequences:

$$\{\cdot\} \longrightarrow \{\Lambda_1\} \longrightarrow \{\Lambda_1,\Lambda_2\} \longrightarrow \text{etc.}; \qquad (4a)$$

$$p(\theta) \longrightarrow p(\theta|\Lambda_1) \longrightarrow p(\theta|\Lambda_1,\Lambda_2) \longrightarrow \text{etc.}; \qquad (4b)$$

$$E[Z] \longrightarrow E[Z|\Lambda_1] \longrightarrow E[Z|\Lambda_1,\Lambda_2] \longrightarrow \text{etc.}, \qquad (4c)$$

In the most general case a model for the learning process can become complex. The computations to be performed at any time may depend on the entire set of priori observations, as is shown by the general form of Bayes' rule

$$p(\theta|\Lambda_1, \ldots \Lambda_n) = \frac{p(\Lambda_n|\theta,\Lambda_1, \ldots \Lambda_{n-1})\, p(\theta|\Lambda_1, \ldots \Lambda_{n-1})}{\int p(\Lambda_n|\theta,\Lambda_1, \ldots \Lambda_{n-1})\, p(\theta|\Lambda_1, \ldots \Lambda_{n-1})\, d\theta} \tag{5}$$

Equation (5) indicates how the new probability density for θ can be computed from the old density; but the computation requires that the probability of Λ_n be known as a function of θ and of all the previous observations, i.e., as $p(\Lambda_n|\theta,\Lambda_1, \ldots \Lambda_{n-1})$. It is often possible to simplify this computation, however. If it be assumed that the different sets of learning observations are conditionally independent (of each other) given θ,* Eq. (5) can be simplified to

$$p(\theta|\Lambda_1, \ldots \Lambda_n) = \frac{p(\Lambda_n|\theta)\, p(\theta|\Lambda_1, \ldots \Lambda_{n-1})}{\int p(\Lambda_n|\theta)\, p(\theta|\Lambda_1, \ldots \Lambda_{n-1})\, d\theta} \tag{6}$$

---

*With this assumption of conditional independence, for any two different sets Λ_i and Λ_j,

$$p(\Lambda_i,\Lambda_j) = \int p(\Lambda_i,\Lambda_j|\theta)\, p(\theta)\, d\theta = \int p(\Lambda_i|\theta)\, p(\Lambda_j|\theta)\, p(\theta)\, d\theta$$

while

$$p(\Lambda_i)\, p(\Lambda_j) = \int\int p(\Lambda_i|\theta)\, p(\theta)\, p(\Lambda_j|\phi)\, p(\phi)\, d\theta d\phi$$

Comparing these two equations it is seen that in general

$$p(\Lambda_i,\Lambda_j) \neq p(\Lambda_i)\, p(\Lambda_j)$$

If $p(\theta)$ is a delta function, however, the inequality becomes an equality. Thus, this conditional-independence assumption may be interpreted as an assumption that, if θ were known, the Λ_i would be statistically independent of each other. With θ unknown, however, the statistical dependence of each Λ_i on θ introduces interdependence among the Λ_i themselves. This interdependence among the Λ_i makes the learning process possible; the interdependence insures that statistical information relating to the value of θ is available in the learning observations.

wherein:

$p(\theta|\Lambda_1, \ldots \Lambda_n)$ = a posteriori probability density of θ, evaluated in the light of the learning observations $\Lambda_1, \ldots \Lambda_n$

$p(\Lambda_n|\theta)$ = likelihood function on θ given by the n^{th} set of learning observations

$p(\theta|\Lambda_1, \ldots \Lambda_{n-1})$ = probability density of θ, evaluated in the light of $\Lambda_1, \ldots \Lambda_{n-1}$

= a posteriori probability density after n-1 sets of learning observations

= a priori probability density just prior to taking n^{th} set of learning observations.

Expanding Eq. (3) to include the improved density calculated from Eq. (6), results in

$$E[Z|\Lambda_1, \ldots \Lambda_n] = \int E[Z|\theta]\, p(\theta|\Lambda_1, \ldots \Lambda_n)\, d\theta \qquad (7)$$

wherein $E[Z|\theta]$ is assumed independent of $\Lambda_1, \ldots \Lambda_n$.

D. IMPLEMENTATION OF LEARNING MODEL

The learning process indicated by Eq. (7) can be implemented as shown in Figure 1.* The process is reiterative, with the same computations performed after obtaining each set of learning observations, but with the probability density on θ updated each time it is used in the computation.

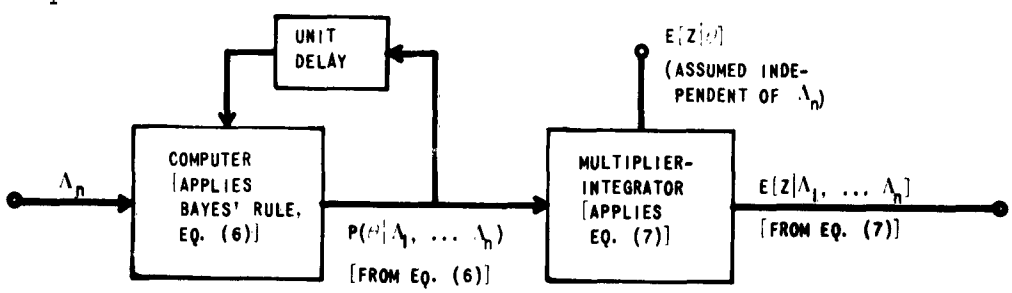

FIG. 1. MODEL FOR LEARNING PROCESS.

*For a model applicable in the more general case, where the conditional-independence assumption is not involved, see Chapter V, Section F.

The special case covered by Eqs. (6) and (7) and Fig. 1, though subject to limitations because of the assumption of the conditional independence of the learning observations $\Lambda_1, \ldots \Lambda_n$, is an important one; in fact, it is the case of primary interest in this investigation. Many of the results of the study are valid for more general cases, however; hence, in the development of the theory of the learning process the possibility of more general results is indicated.

E. DISCUSSION OF LEARNING MODEL

The learning model proposed herein is only one of many possible models. Before it is analyzed in detail some of the implications of the model should be discussed.

In proposing the model a Bayesian approach to the learning problem is used. This approach is often criticized as relying too much on subjective information, especially in the choice of a priori probability distributions. A priori information is seldom exact, so that the a priori probability distributions are normally fairly arbitrary.* On the other hand, Bayesian methods usually allow the use of all available a priori information, even if some subjective elements are involved. Such methods are often applied in cases where the information available is subjective; yet these methods have been found to give reasonable results. A detailed discussion of the implications of the Bayesian approach is given by Savage [Ref. 6].

The model analyzed in the present investigation can also be considered to be a decision-theory model. The methods of statistical decision theory (a theory that has been developed largely on Bayesian lines) normally involve assuming a priori probability distributions, performing experiments to obtain additional information, then making the type of computations indicated in Fig. 1.

If the model illustrated in Fig. 1 is considered as a model of a statistical-decision-theory computation, the techniques of decision

---

*One of the most important results of the work reported here is to indicate reasonable methods for choosing a priori probability functions. The methods, though rational, do not remove the subjective element from the a priori judgment, however.

theory can be used to optimize the performance of a physical or other system under consideration. At least, the performance will be optimum if the correct assumptions are made in the analysis. Since, as noted above, some of these assumptions are almost always subjective, the form of the "optimum" system found by one person may differ from that obtained by another. It can be said that, if the assumptions made by a particular investigator for the analysis are the best that his knowledge allows him to make, then the system performance is, to the extent of his knowledge, optimum; but claims stronger than this are not defensible.* As the number of learning observations increases, however, the subjective elements become relatively unimportant, since the a posteriori probability distributions** become largely independent of the a priori distributions [Ref. 5].

A characteristic of the Bayesian approach that distinguishes it from most other approaches to the learning problem is the fact that no specific value of the unknown parameter θ is selected at any one time. Rather, a probability distribution $p(\theta)$ over the possible values of θ is always considered, and the expectation of the performance criterion is computed based on this distribution [see Eqs. (1), (3), and (7)]. Another approach to the problem would be to estimate a specific value of θ in some way, then to use the estimate as if it were the true value of θ. The two approaches are normally equivalent in the limit as the number of learning observations increases without limit. The common estimates of θ (for example, maximum-likelihood estimates or Bayes estimates) converge in the limit to the true value of the parameter, this convergence taking place with probability one. Similarly, it will be shown that the probability density function $p(\theta|\Lambda_1, \ldots \Lambda_n)$ obtained in the learning-process model developed in this paper converges with probability one to a delta function at the true value of θ. Except for this limiting case, however, specific values of θ are not selected,

*This interpretation is similar to the "personalistic" interpretation of probability theory advocated by Savage [Ref. 6].

**I.e., the probability distributions obtained at the end of the entire sequence of computations.

although the probability densities discussed in connection with the learning model would probably be useful in arriving at a specific estimate of θ.

The significance of the use of a probability distribution $p(\theta)$ over the possible values of θ deserves some comment. A number of interpretations of the significance of this distribution are possible. For example, θ could be considered to be chosen from an ensemble of possible values according to the probability density $p(\theta)$; or the assumption might be made that the uncertainty about θ is caused by some noise (i.e., irrelevant interference) in the selection process. Or, without any explanation at all, it may simply be considered that θ is a random variable representing available knowledge of the unknown parameter. The result of the procedure is probably more important than its justification. The essential point, no matter how interpreted, is that the parameter θ is basically to be treated as a random variable.

III. THE LEARNING PROCESS AND PROBABILITY DISTRIBUTIONS

A. EARLY STUDIES OF THE LEARNING PROCESS

Earlier investigators have analyzed a number of examples and special cases of the learning process [Refs. 7-17]. Some of these earlier investigations furnished the impetus for developing the more general learning model proposed in the present paper. Examples of special interest are those that fall within the special case covered by Eq. (7) and Fig. 1, wherein the learning observations are assumed conditionally independent given θ. Important examples of the learning process involve the application of learning techniques to the pattern-recognition problem. The analysis of the pattern-recognition problem, per se, is only of peripheral interest at this point, but the problem does present an interesting challenge to the learning technique. Therefore, enough of the theory of the pattern-recognition problem will be developed to show that the learning model illustrated in Fig. 1 is applicable (with minor, theoretically insignificant, modifications).

B. THE PATTERN RECOGNITION PROBLEM

It is assumed that there exist r possible patterns, designated by the indices $1, 2, \ldots r$, and that it is desired to classify an observation X as representing one of these patterns. The criterion of excellence Z is taken as the fraction of the patterns identified correctly. Thus, $E[Z]$ is the probability of correct identification.*

Clearly, $E[Z]$ can be maximized by maximizing its value for any given observation. That is, for any given X, the conditional expectation $E[Z|X]$ is to be maximized. But $E[Z|X]$ is the conditional probability of correct identification given the observation X and hence is maximized if the pattern with highest probability of being correct is chosen. Putting these requirements together, it is found that the optimum strategy, or the strategy with maximum probability of correct

*This is the criterion obtained with a statistical-decision-theory approach and a zero-one loss function (i.e., zero loss for a correct decision, loss of one unit for an error).

identification, is to pick the pattern for which the conditional probability $P(i|X)$ is maximum. This strategy can be implemented by computing $P(i|X)$ for each i, or pattern class, then feeding the results of these computations into a comparator that selects the class for which $P(i|X)$ is maximum. This leads to the implementation shown in Fig. 2.

A few modifications of the procedure indicated in Fig. 2 are normally made in implementing such a system. Expanding $P(i|X)$ by Bayes' rule:

$$P(i|X) = \frac{p(X|i)\,P(i)}{p(X)} \qquad (8)$$

where:

$P(i|X)$ = a posteriori probability of the i^{th} pattern class given the observation X [this function is assumed to be evaluated in the light of X by Eq. (8)]

$p(X|i)$ = conditional probability density of the observation X given that the i^{th} pattern is being observed (this density is assumed known as a function of X for any pattern class--at least, in the conventional pattern recognition problem being discussed at this point it is known)

$P(i)$ = a priori probability of the i^{th} pattern class (this probability is also assumed known for each pattern class in the conventional problem)

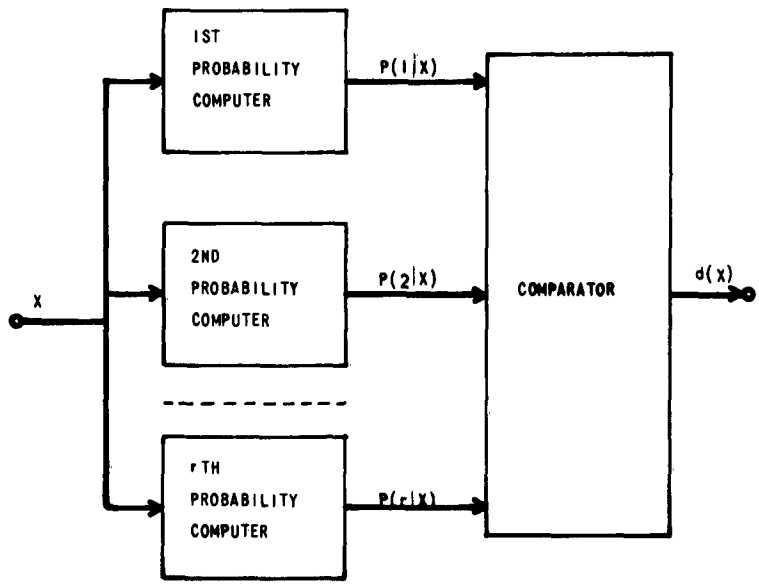

FIG. 2. PATTERN-RECOGNITION SYSTEM.

$p(X)$ = unconditional probability density of the observation X (the availability of this density is unimportant as the discussion below shows that it is not actually needed).

Since $p(X)$ does not depend on i it can be discarded as a variable and attention can be focused on maximizing $p(X|i)P(i)$. It is further assumed (for simplicity) that all $P(i)$'s are known and equal, so that all that remains is merely to maximize $p(X|i)$.

The earlier work on the pattern-recognition problem [Refs. 7-10] has been based on the computation of $p(X|i)$ when some parameter θ_i in this probability-density function is unknown. The basic equations are slight modifications of Eqs. (1) and (7).*

$$p(X|i) = \int p(X|i,\theta_i)p(\theta_i)\, d\theta_i \tag{9}$$

$$p(X|i,\Lambda_{i1}, \ldots \Lambda_{in}) = \int p(X|i,\theta_i)p(\theta_i|\Lambda_{i1}, \ldots \Lambda_{in})\, d\theta_i \tag{10}$$

The Λ_{ij} are assumed to be sets of learning observations from the i^{th} pattern class.

Since the procedure for all pattern classes is identical, the subscripts i are now dropped to simplify notation.

C. OTHER EXAMPLES OF THE LEARNING PROCESS

Abramson and Braverman [Refs. 7-9] have been primarily concerned with the case where $p(\mathbf{X})$ is known to be Gaussian, $p(\mathbf{X}) \sim N(\mathbf{M}, \mathbf{K})$,**

*It would be simple to make the correspondence between Eqs. (1) and (9) and between Eqs. (7) and (10) more exact by defining random variables with expectations $p(X|i)$ and $p(X|i,\Lambda_{i1}, \ldots \Lambda_{in})$.

**Symbols that represent matrices (including vectors) are in boldface type. When a symbol is used to represent a variable that could be either a real number or a vector or matrix (for example, the general parameter θ), ordinary type is used, however. The notation $p(\mathbf{X}) \sim N(\mathbf{M}, \mathbf{K})$ may be read, "The probability density of the vector $\mathbf{X}$ (actually the joint density of the components of $\mathbf{X}$) is normally distributed (or Gaussian) with mean vector $\mathbf{M}$ and covariance matrix $\mathbf{K}$."

with the covariance matrix $\mathbf{K}$ known but the mean vector $\mathbf{M}$ unknown. In other words, Abramson and Braverman's unknown parameter θ is the mean vector $\mathbf{M}$ of a Gaussian density. They assume a Gaussian a priori density for $\mathbf{M}$, $p(\mathbf{M}) \sim N(\mu_0, \phi_0)$ and obtain an a posteriori density, $p(\mathbf{M}|\Lambda_1)$, which is also Gaussian, $p(\mathbf{M}|\Lambda_1) \sim N(\mu_1, \phi_1)$ with μ_1 and ϕ_1 easily computed from μ_0, ϕ_0, and Λ_1. The densities for $\mathbf{X}$, both a priori and a posteriori, are also Gaussian, $p(\mathbf{X}) \sim N(\mu_0, \phi_0 + \mathbf{K})$ and $p(\mathbf{X}|\Lambda_1) \sim N(\mu_1, \phi_1 + \mathbf{K})$.

The second stage in the learning process under study illustrates why this particular process is feasible. Since $p(\mathbf{M}|\Lambda_1)$ is of the same form as $p(\mathbf{M})$ (i.e., Gaussian), and the second stage involves the same computations as the first stage with $p(\mathbf{M}|\Lambda_1)$ substituted for $p(\mathbf{M})$, Gaussian probability densities are again obtained for $\mathbf{M}$ and $\mathbf{X}$. By induction it is seen that this will happen after each set of learning observations. Hence, the form of the learning system remains fixed as more learning observations are taken.

After each set of learning observations Λ_n, the new mean μ_n for the density on $\mathbf{M}$ is computed as a weighted average of μ_{n-1} and the average of the observations in Λ_n. In the limit, as the number of learning observations approaches infinity, μ_n approaches the average of all the learning observations. It is known, from the strong law of large numbers [Ref. 18], that with probability one the average of the observations approaches the true value $\mathbf{M}_o$ of the mean. At the same time, the elements of the covariance matrix ϕ_n approach zero. Thus, the limiting form of $p(\mathbf{M}|\Lambda_1, \ldots \Lambda_n)$ is $N(\mathbf{M}_o, \mathbf{0})$. Comparing this with the multivariate Dirac delta function, it is found that the limiting form of the a posteriori density on $\mathbf{M}$ is a Dirac delta function at the true value of the mean.

If this delta function is put into the equation for $p(\mathbf{X}|\Lambda_1, \ldots \Lambda_n)$, it is found that the density approaches the form for known parameters. Hence, the entire system converges to the form it would take if the parameters were known.

The solution for the problem of learning the unknown mean was obtained in a fairly simple manner. The assumption of a Gaussian a priori

probability density on $\mathbf{M}$ is the obvious assumption to make since $\mathbf{M}$ is a parameter in a Gaussian density. This assumption gives Gaussian a posteriori densities on $\mathbf{M}$, and insures that all the densities required are Gaussian.

Keehn [Ref. 10] has analyzed a similar problem and obtained similar results. For his problem the assumptions that keep the form of the learning system fixed are less obvious, however. Keehn has analyzed the problem of learning the covariance matrix $\mathbf{K}$ for a Gaussian density when the mean vector $\mathbf{M}$ is known.

The key assumption necessary to solve the unknown covariance problem is the assumption of a Wishart a priori density over the elements of the inverse covariance matrix $\mathbf{K}^{-1}$.* The a posteriori density on the elements of $\mathbf{K}^{-1}$ is also Wishart, with new parameters calculated from the old parameters and the learning observations. The limiting form of the a posteriori density is again a delta function at the true values of the unknown parameters, in this case the true values of the components of the inverse covariance matrix.

The probability density for $\mathbf{X}$ turns out, in this case, to be a Student density instead of the Gaussian density one might expect. As the number of learning observations approaches infinity, however, the limiting form of the Student density becomes Gaussian with the true mean vector and covariance matrix. Hence, the limiting form of the a posteriori density on $\mathbf{X}$ is as desired.

Keehn has analyzed in a similar manner the case where both $\mathbf{K}$ and $\mathbf{M}$ are unknown. He obtained analogous results by assuming a composite Wishart-Gaussian density on the elements of $\mathbf{K}^{-1}$, $\mathbf{M}$.** The a posteriori density is also of this composite form and converges to a delta function at the true values of the unknown parameters. The density on $\mathbf{X}$ is a modified form of the Student density, which approaches the true Gaussian density.

---

*The form of this density is given in Chapter VI, Table 2, Case 6.

**The form of this density is given in the Appendix, Eq. (A-7).

D. FEASIBILITY OF THE LEARNING PROCESS AS DETERMINED BY PROBABILITY DISTRIBUTIONS

The examples cited above illustrate one method of guaranteeing that the learning process is feasible. If it is possible to pick an a priori density $p(\theta)$ for θ such that the a posteriori density $p(\theta|\Lambda_1, \ldots\Lambda_n)$ is of the same form (e.g., both Gaussian or both Wishart), then the Bayes' rule computer merely computes new values for the parameters describing the density on θ in terms of the old values and the learning observations. If the form of the density is preserved after one set of learning observations, the arguments used for the Gaussian case show that it is preserved no matter how many learning observations are taken. Hence, the learning process is feasible in the sense under consideration--i.e., in the sense that a fixed form of computations is applicable throughout the entire process.

The learning process is considered to be feasible if the computations necessary after taking learning observations are fixed, neither the number nor the forms of the computations changing. This requirement of a fixed set of computations is imposed from the point of view of engineering feasibility. If the system can learn by performing a fixed set of computations after each observation period, the engineering problems in designing an actual system may be soluble; if the system has to be reprogrammed periodically, or if the number of computations necessary grows without bound, the design problems almost certainly are not soluble.

1. Reproducing-Type Distributions

In the present investigation, probability distributions that preserve their form under Bayes' rule, i.e., for which the a priori and a posteriori distributions have the same form, will be designated as "reproducing-type distributions." Besides the investigators mentioned above, a number of other persons have utilized distributions of this type. Bellman [Ref. 11] has utilized a beta density for learning the parameter characterizing a binomial distribution; Mosimann [Ref. 12] has utilized the "multivariate beta" or Dirichlet distribution for the parameters of a multinomial distribution; Turin [Ref. 13] has used the "generalized Rayleigh" or Rician density for learning the amplitude

and phase characteristics of a radio channel; and Kailath [Ref. 14] has utilized a Gaussian distribution for learning the unknown mean of a Gaussian distribution, obtaining results similar to those of Abramson and Braverman in a different manner. None of these workers give methods for finding reproducing-type distributions, however. The only general method of finding reproducing-type distributions that has been found in the literature is that of Raiffa and Schlaifer [Ref. 15]. These authors discuss an important class of reproducing-type distributions--a class that includes all the reproducing distributions mentioned above save the Rician distribution utilized by Turin.*

2. Nonreproducing Distributions

Lest the reader gain the impression that reproducing-type distributions always exist, so that the problem is merely one of finding the appropriate reproducing distribution, attention is called to one example of a case where no reproducing distributions exist. This example is taken from a problem studied by Daly [Refs. 16 and 17], which is similar to the problems studied by Abramson, Braverman, et al. The chief difference between Daly's problem and the cases hitherto mentioned lies in the form of the information given to the learning system during the learning process. An important assumption in the analysis of the examples previously considered has been the assumption that the learning observations were classified--i.e., the system was told to which pattern each learning observation corresponded. This assumption made it possible to state that the Λ_{ij} in Eq. (10) consisted of samples from the i^{th} pattern class.** Daly assumed that the system was not given this

*The forms of all these densities, including that used by Turin, are derived in Chapter VI and in the Appendix.

**In a typical application of this theory the system would be given a set of classified patterns during a training period, then would be told to identify unclassified patterns later. In a few cases the correct classification of patterns might be available with a slight delay, with a decision needed earlier. The same techniques could be used as in the first case, but with the added possibility of indefinitely continuing the training period.

information, either during the learning process or during the recognition process. The two problems may be distinguished by calling the former the "perfect-teacher" problem and the latter the "no-teacher" problem.

A simple example of the "no-teacher" problem would allow for two alternative hypotheses: either (1) both noise and a one-dimensional signal of unknown magnitude m are present; or (2) the noise alone is present. Assuming Gaussian noise distribution with zero mean and variance σ^2, and assuming also that the two hypotheses are equally probable, the conditional probability density of an observation X given m is:

$$p(X|m) = \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}\,\sigma} \left\{ \exp\ -(X-m)^2/2\sigma^2 + \exp\ -X^2/2\sigma^2 \right\} \quad (11)$$

If an a priori probability density $p(m)$ is assumed and if a set $X_1, \ldots X_n$ of measurements chosen according to the density given by Eq. (11) are used as learning observations, it is found that

$$p(m|X_1, \ldots X_n) = \frac{p(X_1, \ldots X_n|m)\, p(m)}{\int p(X_1, \ldots X_n|m)\, p(m)dm}$$

$$= \frac{\prod_{i=1}^{n} \left\{ \exp\ -(X_i-m)^2/2\sigma^2 + \exp\ -X_i^2/2\sigma^2 \right\} p(m)}{\int \prod_{i=1}^{n} \left\{ \exp\ -(X_i-m)^2/2\sigma^2 + \exp\ -X_i^2/2\sigma^2 \right\} p(m)dm} \quad (12)$$

In each of the earlier examples the a posteriori density $p(\theta|X_1, \ldots X_n)$ was expressible in terms of a fixed number of parameters no matter how many learning observations were taken. Thus, the form of the density did not change as the learning observations progressed. In the case of learning a Gaussian mean $\mathbf{M}$, only two parameters, μ_n and ϕ_n, were necessary. Since the Wishart density is expressed in terms of a fixed set of parameters, a similar situation was true for

learning the covariance matrix $\mathbf{K}$ or for learning both $\mathbf{M}$ and $\mathbf{K}$. This is not the case with the density in Eq. (12), however. In fact, no nondegenerate form for $p(m)$ has been found that allows $p(m|X_1, \ldots, X_n)$ to be expressed in terms of fewer than n parameters (one for each X_i). It is shown in Chapter VI, Section D, that expression in terms of fewer than n parameters is impossible with any nondegenerate $p(m)$; hence, the form of the density keeps changing as long as the learning observations are continued.

The example of the "no-teacher" problem clarifies what is meant by saying that the a priori and a posteriori densities are of the same form; this requirement must be interpreted to include expression of the densities in terms of a fixed number of parameters. Otherwise, the density in Eq. (12) might conceivably be considered to be reproducing, since the expression in the last part of this equation is always valid. The example also indicates that it cannot automatically be assumed in any particular case that reproducing-type densities exist.

E. PROBLEMS FOR FURTHER INVESTIGATION

Examples of the learning process studied in this chapter have described three main problems:

1. To find general conditions under which the a posteriori probability density approaches a delta function at the true value of the unknown parameter.
2. To find conditions guaranteeing the existence of reproducing-type probability distributions.
3. To find the forms of any reproducing-type probability distributions that may exist in a particular case.

These problems are investigated in the following chapters.

IV. CONDITIONS UNDER WHICH THE A POSTERIORI DISTRIBUTION APPROACHES A DELTA FUNCTION

This chapter considers the first problem posed at the end of Chapter III: to find general conditions under which the a posteriori probability distribution approaches a delta function at the true value of the unknown parameter.

A. THE CONVERGENCE THEOREM

In each of the examples of learning processes discussed in Chapter III the limiting form of the a posteriori density $p(\theta|\Lambda_1, \ldots \Lambda_n)$ as n increases is a delta function at the true value of θ. The conditions needed to insure that this is so are simple: it must be possible to calculate the true value of θ from an infinite sequence of observations, and this true value must not be ruled out by $p(\theta)$, the a priori probability distribution on θ. More rigorously:

Theorem I. Assume that the following conditions are satisfied:

1. θ_o is the true value of θ
2. The a priori density $p(\theta) > 0$ in some sphere containing θ_o
3. The a posteriori densities $p(\theta|\Lambda_1, \ldots \Lambda_n)$ are calculated by Bayes' rule
4. There exists a sequence of functions $f_n(\Lambda_1, \ldots \Lambda_n)$ converging to θ_o with probability one.

Then $p(\theta|\Lambda_1, \ldots \Lambda_n) \to \delta(\theta - \theta_o)$ with probability one, where $\delta(\theta - \theta_o)$ is a Dirac delta function (of the same dimension as θ).

Proof: Theorem I is an immediate consequence of the zero-one law of probability theory as stated by Loéve [Ref. 18, p. 398]. The statement of this law used here is, "The sequence $P(B|Y_1, \ldots Y_n)$ of conditional probabilities of a property B of the sequence $Y_1, Y_2, \ldots$ given the first n terms of the sequence converges almost surely to 1 or 0 according as the sequence has or has not this property." If B is a sphere in the range of θ, then the event that

$$\theta_o = \lim_{n\to\infty} f_n(\Lambda_1, \ldots \Lambda_n) \in B^*$$

is an event defined on the Λ_i and hence satisfies Loeve's definition of a "property" of the sequence. Therefore,

$$P(B|\Lambda_1, \ldots \Lambda_n) = \int_B p(\theta|\Lambda_1, \ldots \Lambda_n)\, d\theta \longrightarrow 1 \text{ or } 0 \qquad (13)$$

according as θ_o is or is not in B. Equation (13) is equivalent to the statement that $p(\theta|\Lambda_1, \ldots \Lambda_n)$ converges to $\delta(\theta - \theta_o)$.**

Since Theorem I and its proof are fairly abstract, the significance of the assumptions should be pointed out. Assumption (4) guarantees that the event that $\theta_o \in B$ is a property of the sequence. Assumption (1) guarantees that this event is true, or that the sequence has the desired property. Assumption (3) guarantees that the correct forms are used for the a posteriori probabilities, since these probabilities are calculated by the standard methods of probability theory. The other assumption, number (2), is hidden in Loéve's statement of the zero-one law. In all of the material he treats, Loéve assumes the events considered have positive probability. Assumption (2) insures that this is true.

From the definition of the Dirac delta function and Eq. (3) there is derived the important

Corollary: If the assumptions in Theorem I are satisfied, $E[Z|\Lambda_1, \ldots \Lambda_n] \to E[Z|\theta_o]$ with probability one, where Z is a random variable representing a selected performance criterion.

---

*The symbol $\in$ in this equation should be read "is in" or "belongs to."

**Theorem I is based on Theorem 5.1 of Braverman [Ref. 7, p. 29]. The material just presented comprises a more precise statement of the theorem and simplifies the proof. The proof is still quite abstract, however, despite its deceptively simple appearance. Those readers unable to follow the proof completely may treat it as a plausibility argument.

This corollary indicates that the entire system approaches the form it would take if θ_o were known to be the true value of θ.

B. DISCUSSION OF THEOREM

Theorem I is more general in its import than may at first be apparent. No statements have been made as to whether a "teacher" is present or not. It has not been required that any type of independence hold, nor does Loeve require independence for his theorem. It is merely required that the sequence of functions $f_n(\Lambda_1, \ldots \Lambda_n)$ exist. Such a sequence can exist either with or without a teacher, either with or without independence.

The requirement that this sequence of functions exist is simply a method of saying that the true value of θ must, with probability one, be determinable from an infinite sequence of learning observations. If it be assumed that the sets of learning observations consist of single observations, i.e., $\Lambda_i = \{X_i\}$, and that the X_i are conditionally independent given θ (the same independence assumption used in Chapter II), this requirement can be put into a more easily visualized form. In this case if a function of a single observation, $f(X_i)$, such that

$$E[f(X_i)|\theta] = \theta, \tag{14}$$

exists, then by the strong law of large numbers,

$$\frac{1}{n} \sum_{i=1}^{n} f(X_i) \rightarrow \theta_o \tag{15}$$

with probability one, where θ_o is the true value of θ.*

*In applying the strong law of large numbers to this case, it is necessary to recall the earlier interpretation of the requirement that the X_i be independent given θ. In Chapter II, this requirement was interpreted to mean that if θ were known the X_i would be independent. The knowledge available about θ does not affect the convergence of Eq. (15); so the strong law is applied as if it were known that θ equals θ_o.

As an example, in the case of the unknown mean of a Gaussian distribution, the sample average

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i$$

converges to the true value of the mean with probability one. Similarly, for the case of an unknown covariance matrix, the sample covariance matrix

$$\frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \mathbf{M})_t (\mathbf{X}_i - \mathbf{M})$$

converges to the true covariance matrix with probability one.

Theorem I can also be applied to the simple example of the "no-teacher" problem discussed in Chapter III. For the density given by Eq. (11),

$$E[X|m] = \frac{1}{2} m. \tag{16}$$

Hence, by Eqs. (14) and (15)

$$\frac{2}{n} \sum_{i=1}^{n} X_i \longrightarrow m_o \tag{17}$$

with probability one, where m_o is the true value of m. This result agrees with Daly's application of limiting arguments [Refs. 16 and 17] to show that the limiting form of the optimum system is the form it would take if m were known.

As the conditions of Theorem I are met for most probability distributions of practical significance, this theorem provides reasonably general conditions insuring that the limiting form of the a posteriori density is a delta function at the true value of θ. Thus, Theorem I affords a solution to the first of the three problems posed at the end of Chapter III.

C. ILLUSTRATION OF CONVERGENCE

An illustration of the manner in which the a posteriori density approaches a delta function is given by Fig. 3. In this figure are plotted probability densities for the parameter P characterizing a binomial distribution. A uniform a priori density over the interval from 0 to 1 has been assumed, and the a posteriori density $p(P|\Lambda_1, \ldots \Lambda_n)$ has been plotted under the assumption that equal numbers (1, 2, 4, 8 and 16) of occurrences of each of the two possible

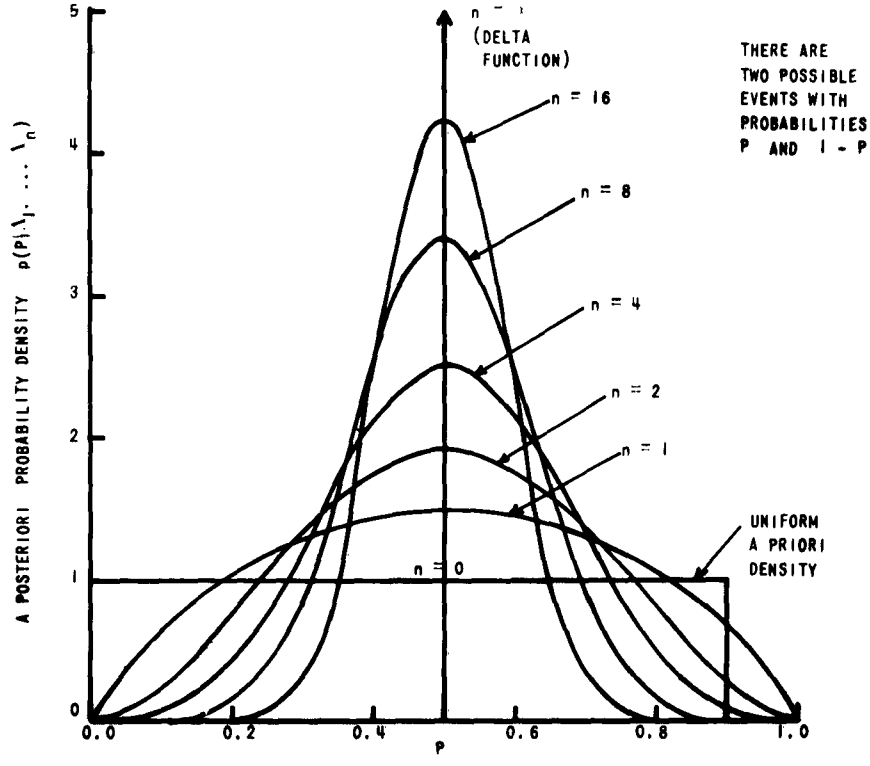

FIG. 3. PROBABILITY DENSITIES FOR THE PARAMETER P CHARACTERIZING A BINOMIAL DISTRIBUTION.

events have been observed.* The conclusion from the plot is that the value of P becomes known more and more accurately as more observations are taken--this is illustrated by the continuously decreasing width of the plots in Fig. 3--with the true value of P becoming known exactly after an infinite number of observations, when the density becomes a delta function at the true value of P, P = 1/2.

*Since the a priori density p(P) is uniform and none of the a posteriori densities are uniform (in this case all of the a posteriori densities are beta), the a priori density in this example is not reproducing-type. However, since all the a posteriori densities are of the same form, the densities may be considered to become reproducing-type after one observation. It is shown in Chapter V, Section D, that a posteriori densities often become reproducing-type after a few observations even when the a priori density is not reproducing-type.

V. CONDITIONS FOR REPRODUCING-TYPE PROBABILITY DISTRIBUTIONS

This chapter attacks the second and third problems posed at the end of Chapter III: namely, the problem of finding conditions guaranteeing the existence of reproducing-type probability distributions, and also the problem of finding the forms of any such distributions that may exist.

A reproducing-type probability distribution has been defined as one in which the a posteriori distribution $p(\theta|\Lambda_1, \ldots \Lambda_n)$ has the same form as the distribution $p(\theta)$ assumed a priori, the two distributions being related through Bayes' rule applied in the light of a series of learning observations $\Lambda_1, \ldots \Lambda_n$ (Eqs. (2) and (6)). The first step in the present study, therefore, is to find a convenient method for analyzing the form of $p(\theta|\Lambda_1, \ldots \Lambda_n)$ in any particular case.

A. FACTORIZATION OF A POSTERIORI DENSITY (ASSUMING LEARNING OBSERVATIONS ARE CONDITIONALLY INDEPENDENT GIVEN θ)

A principal difficulty in analyzing the form of the a posteriori probability density $p(\theta|\Lambda_1, \ldots \Lambda_n)$ as it is given by Bayes' rule arises from the arbitrary nature of the a priori density $p(\theta)$. The only real requirement put on the a priori density is that it be a true probability density; hence, it must be non-negative and integrate to one. Since $p(\theta)$ is involved in the computation of each of the a posteriori densities $p(\theta|\Lambda_1, \ldots \Lambda_n)$, this introduces some arbitrariness into each of these a posteriori densities. This may be illustrated by writing Bayes' rule in terms of the likelihood of the complete sequence of sets of learning observations, i.e., in terms of $p(\Lambda_1, \ldots \Lambda_n|\theta)$:

$$p(\theta|\Lambda_1, \ldots \Lambda_n) = \frac{p(\Lambda_1, \ldots \Lambda_n|\theta)\, p(\theta)}{\int p(\Lambda_1, \ldots \Lambda_n|\theta)\, p(\theta)\, d\theta} . \qquad (18)$$

Fortunately, the expression in Eq. (18) for the a posteriori density may be factored in a manner that simplifies analysis of its form.

Theorem II: Assume the likelihood $p(\Lambda_1, \ldots \Lambda_n|\theta)$ is greater than zero and is an integrable function of θ. Then $p(\theta|\Lambda_1, \ldots \Lambda_n)$ can be expressed as

$$p(\theta|\Lambda_1, \ldots \Lambda_n) = \hat{p}(\theta|\Lambda_1, \ldots \Lambda_n) \cdot \frac{p(\theta)}{\hat{E}[p(\theta)|\Lambda_1, \ldots \Lambda_n]} \tag{19}$$

where

$$\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n) = \frac{p(\Lambda_1, \ldots \Lambda_n|\theta)}{\int p(\Lambda_1, \ldots \Lambda_n|\theta)\, d\theta} \tag{20}$$

is a probability density on θ depending only on the observations, and where $\hat{E}[p(\theta)|\Lambda_1, \ldots \Lambda_n]$ is the expectation of the a priori density $p(\theta)$ taken with respect to the density $\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n)$. Further, if $p(\theta)$ is bounded and $p(\theta_o) > 0$, then

$$p(\theta|\Lambda_1, \ldots \Lambda_n) \rightarrow \delta(\theta - \theta_o) \tag{21}$$

with probability one if and only if

$$\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n) \rightarrow \delta(\theta - \theta_o) \tag{22}$$

Proof: The function $\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n)$ is by its definition in Eq. (20) a legitimate and well-defined probability density, since it has been assumed that $p(\Lambda_1, \ldots \Lambda_n|\theta) > 0$ and is integrable. Rewriting Eq. (18) in the form

$$p(\theta|\Lambda_1, \ldots \Lambda_n) = \frac{p(\Lambda_1, \ldots \Lambda_n|\theta)}{\int p(\Lambda_1, \ldots \Lambda_n|\theta)\, d\theta} \cdot \frac{p(\theta)}{\int \left[\frac{p(\Lambda_1, \ldots \Lambda_n|\theta)}{\int p(\Lambda_1, \ldots \Lambda_n|\theta)\, d\theta}\right] p(\theta) d\theta} \tag{18a}$$

and incorporating the definition of $\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n)$ in Eq. (20) it is seen that $p(\theta|\Lambda_1, \ldots \Lambda_n)$ may be written in the form in Eq. (19).

To prove the convergence portions of Theorem II, assume $\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n) \to \delta(\theta - \theta_o)$ as specified in Eq. (22). Then, since $p(\theta_o) > 0$ by assumption, and $\hat{E}[p(\theta)|\Lambda_1, \ldots \Lambda_n]$ approaches $p(\theta_o)$ as $\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n) \to \delta(\theta - \theta_o)$,

$$p(\theta|\Lambda_1, \ldots \Lambda_n) \to \frac{p(\theta)}{p(\theta_o)} \cdot \delta(\theta - \theta_o) = \delta(\theta - \theta_o) \tag{23}$$

Conversely, if it be assumed that $p(\theta|\Lambda_1, \ldots \Lambda_n) \to \delta(\theta - \theta_o)$, then Eq. (19) indicates that

$$\begin{aligned} \hat{p}(\theta|\Lambda_1, \ldots \Lambda_n) &= p(\theta|\Lambda_1, \ldots \Lambda_n)\, \hat{E}[p(\theta)|\Lambda_1, \ldots \Lambda_n] \,/p(\theta) \\ &\to \delta(\theta - \theta_o)\, \hat{E}[p(\theta)|\Lambda_1, \ldots \Lambda_n]/p(\theta) \end{aligned} \tag{24}$$

Since $\hat{E}[p(\theta)|\Lambda_1, \ldots \Lambda_n]$ is a constant and $p(\theta)$ has been assumed to be bounded, Eq. (24) can be valid only if $\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n) \to \delta(\theta - \theta_o)$.

The density $\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n)$, which might be called the "experimental portion" of the a posteriori density, is simply a normalized version of the likelihood. It is a function of $\Lambda_1, \ldots \Lambda_n$ as well as of θ, but it is here assumed that the observations have been made and $\Lambda_1, \ldots \Lambda_n$ have been replaced by the results of the observations. Under these conditions, $\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n)$ is a function of the single variable θ.

The integrability condition on $p(\Lambda_1, \ldots \Lambda_n|\theta)$ in Theorem II is normally fulfilled for large n, as this density tends to become more and more concentrated near the true value of θ, so that the effective

range of integration is small.* In all cases thus far encountered for which the techniques of Theorem II are applicable, $p(\Lambda_1, \ldots \Lambda_n|\theta)$ becomes integrable after a few observations (typically one or two) and remains integrable as more observations are made. Unless otherwise stated, it will henceforth be assumed that this integrability condition is satisfied.

B. EXPERIMENTAL PORTION OF A POSTERIORI DENSITY

Theorem II indicates that, at least after a large number of learning observations, the behavior of $p(\theta|\Lambda_1, \ldots \Lambda_n)$ is primarily determined by the "experimental portion" $\dot{p}(\theta|\Lambda_1, \ldots \Lambda_n)$. Also, the latter density is less arbitrary and consequently easier to work with than is the basic function. The conditions that must be satisfied for the "experimental portion" of the a posteriori density to be reproducing are now to be investigated.

Definition No. 1: The a priori density $p(\theta)$ is said to reproduce itself with respect to the likelihood $p(\Lambda_1|\theta)$ if $p(\theta)$ and the

*Lindley [Ref. 5] has shown that with any reasonably smooth a priori density, the limiting form of the a posteriori density is independent of the a priori density, being Gaussian with means at the maximum likelihood values and with variances decreasing as $1/n$. (Another type of density, possibly a reproducing-type density, may approximate the a posteriori density slightly more accuracy, but both this density and Lindley's Gaussian density approach each other and the delta function limit of Theorem I.) A general proof that the effective range of integration approaches zero is easily deduced from Lindley's result.

The limiting form Lindley obtains is almost identical to the limiting form for the probability density of a maximum-likelihood estimate. This latter density can be found in many standard statistics texts. An alternative approach to proving that the effective range of integration approaches zero could be based on these maximum-likelihood analyses.

An illustration of the manner in which the effective range of integration for $p(\Lambda_1, \ldots \Lambda_n|\theta)$ approaches zero may be deduced from Fig. 3. Since in that figure a uniform a priori density was assumed, the a posteriori density plotted in the figure is proportional to $p(\Lambda_1, \ldots \Lambda_n|\theta)$, and the effective range of integration is the effective width of the plot.

a posteriori density $p(\theta|\Lambda_1)$ are members of the same family of probability densities, differing only in the values of the parameters characterizing densities in this family.

If $p(\theta)$ reproduces itself, the result of the Bayes' rule computation in the learning process is simply to compute new values for the parameters characterizing densities in the family, this computation giving $p(\theta|\Lambda_1)$. The next stage of the learning process involves the same computations save for replacing $p(\theta)$ by $p(\theta|\Lambda_1)$ and using the set Λ_2 of learning observations instead of Λ_1. If these sets of learning observations are of the same type, $p(\theta|\Lambda_1)$ reproduces itself with respect to the likelihood $p(\Lambda_2|\theta)$ if $p(\theta)$ reproduces itself with respect to $p(\Lambda_1|\theta)$. Proceeding by induction, it is seen that $p(\theta|\Lambda_1, \ldots \Lambda_{n-1})$ reproduces itself with respect to $p(\Lambda_n|\theta)$ if $p(\theta)$ reproduces itself with respect to $p(\Lambda_1|\theta)$.

Thus, under the assumed set of conditions, the fact that $p(\theta)$ reproduces itself with respect to the likelihood $p(\Lambda_1|\theta)$ guarantees that all the a posteriori densities are members of the same family of probability densities. At each stage of the learning process the Bayes' rule computer merely computes new values for the parameters describing these densities. The remainder of the computations involved in the learning process, multiplication by $E[Z|\theta]$ and integration, are fixed computations (see Fig. 1) and can always be accomplished in the same manner. Even if the result of this computation cannot be obtained analytically in closed form, it can be obtained by a fixed procedure of numerical integration or by electronic integration. Hence, if $p(\theta)$ reproduces itself with respect to $p(\Lambda_1|\theta)$, the computations necessary for the entire learning process are the same at each stage of the process. It is assured that the system will not have to be reprogrammed in the middle of the learning process.

Strictly speaking, the sets of learning observations or the likelihoods $p(\Lambda_i|\theta)$ should be included in any statement about densities reproducing themselves. In cases where the meaning is clear, however, reference will be made to the densities $p(\theta)$ as being reproducing-type densities, without specific mention of the learning observations.

C. SUFFICIENT STATISTICS

In actual computations of the a posteriori probabilities, it is often unnecessary to have available all the individual learning observations. It often happens that some functions of the learning observations will suffice for computing the a posteriori probabilities. For example, the a posteriori probability density for the mean of a Gaussian distribution given the sample average for the learning observations is the same as the a posteriori density given all the individual observations. A function of the learning observations which, in this sense, contains all the information in the observations relevant to learning θ is called a sufficient statistic for θ.*

In working with sufficient statistics it is considered that they are written in the form of a vector with real-valued components. That is, if $T(\Lambda_1, \ldots \Lambda_n)$ is a sufficient statistic for θ, it is assumed that

$$T(\Lambda_1, \ldots \Lambda_n) = \left(t_1^{(n)}, \ldots t_s^{(n)}\right) \tag{25}$$

where the $t_i^{(n)}$ are real-valued functions of $\Lambda_1, \ldots \Lambda_n$. There follows the obvious

Definition No. 2: The dimension of a sufficient statistic is the number of components in the vector representation of the statistic.

In the case of learning the unknown mean of a Gaussian density mentioned above, the sample average is a sufficient statistic of fixed dimension (d dimensions if a d-variate Gaussian density is being considered). In some cases, however, the only sufficient statistic is equivalent to the learning observations themselves** and no sufficient statistic of fixed dimension exists. The distinction is of fundamental importance, as indicated by Theorem III below.

---

*A general treatment of sufficient statistics has been given by Dynkin [Ref. 19]. Among other things he finds conditions for the existence of sufficient statistics of the forms needed for this study and methods for computing such sufficient statistics.

**The statistic is equivalent to the observations if the observations can be computed from the statistic and vice versa.

It is now possible to state a simple criterion for determining whether the experimental portion of the a posteriori density is reproducing or not. Since this density is not defined before observing Λ_1, the procedure suggested by Definition No. 1 is slightly altered by checking whether $\hat{p}(\theta|\Lambda_1)$ reproduces itself with respect to $p(\Lambda_2|\theta)$ or not.

Theorem III: The probability density $\hat{p}(\theta|\Lambda_1)$ reproduces itself with respect to the likelihood $p(\Lambda_2|\theta)$ if and only if a sufficient statistic for θ of fixed dimension exists.

Proof: To prove this theorem the factorization theorem for sufficient statistics is applied [Ref. 20]. The factorization theorem states that $\left(t_1^{(n)}, \ldots t_s^{(n)}\right)$ is a sufficient statistic for θ if and only if there exist functions f and h such that

$$p(\Lambda_1, \ldots \Lambda_n|\theta) = f\left(t_1^{(n)}, \ldots t_s^{(n)}, \theta\right) h(\Lambda_1, \ldots \Lambda_n) \qquad (26)$$

where f depends on $\Lambda_1, \ldots \Lambda_n$ only through $t_1^{(n)}, \ldots t_s^{(n)}$, and where h does not depend on θ.

Assume a sufficient statistic of fixed dimension exists and let $\left(t_1^{(n)}, \ldots t_s^{(n)}\right)$ be such a sufficient statistic. Then, from Eqs. (20) and 26),

$$\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n) = \frac{f\left(t_1^{(n)}, \ldots t_s^{(n)}, \theta\right)}{\int f\left(t_1^{(n)}, \ldots t_s^{(n)}, \theta\right) d\theta} \qquad (27)$$

This is a fixed function of the parameters $t_1^{(n)}, \ldots t_s^{(n)}$. Hence, the $\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n)$ differ only in the values assigned to these parameters and each reproduces itself with respect to $p(\Lambda_{n+1}|\theta)$.

Conversely, assume $\hat{p}(\theta|\Lambda_1)$ reproduces itself with respect to $p(\Lambda_2|\theta)$. Then there exist r parameters $\alpha_1^{(n)}, \ldots \alpha_r^{(n)}$ and a function g such that

$$\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n) = g\left(\alpha_1^{(n)}, \ldots \alpha_r^{(n)}, \theta\right) \qquad (28)$$

since it is known that all of these densities are of the same form, differing only in the values assigned to parameters. From Eqs. (20) and (28),

$$p(\Lambda_1, \ldots \Lambda_n|\theta) = g\left(\alpha_1^{(n)}, \ldots \alpha_r^{(n)}, \theta\right) \cdot \int p(\Lambda_1, \ldots \Lambda_n|\theta)\, d\theta \tag{29}$$

The last integral is not a function of θ, since this parameter is integrated out of the equation. Hence, by the factorization theorem for sufficient statistics, the α's comprise a sufficient statistic for θ of fixed dimension.

D. REPRODUCING A PRIORI DENSITIES

By combining the results in Theorems II and III, solutions can be obtained to the problems of determining when reproducing-type densities exist and of finding the forms of any that exist.

First, it is noted that the factorization in Eq. (19), Theorem II, expresses $p(\theta|\Lambda_1, \ldots \Lambda_n)$ as the product of $\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n)$ and another function of θ. Hence, if the densities $p(\theta)$, $p(\theta|\Lambda_1)$, ... are all to be of the same form, the densities $\hat{p}(\theta|\Lambda_1)$, $\hat{p}(\theta|\Lambda_1, \Lambda_2)$, ... must all be of the same form. According to Theorem III, this means that a sufficient statistic of fixed dimension must exist.

Second, it may be seen that if $p(\theta)$ is to be a reproducing-type a priori density, it must be of the same form as the a posteriori density $p(\theta|\Lambda_1, \ldots \Lambda_n)$. Hence, $p(\theta)$ must be a function <u>of the form of</u> $\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n)$ multiplied by another function of θ. This condition is stated by postulating that $p(\theta)$ must be of the form

$$\hat{p}(\theta) = \frac{p(\theta|\Lambda_{-m}, \ldots \Lambda_0)\, r(\theta)}{\int p(\theta|\Lambda_{-m}, \ldots \Lambda_0\, r(\theta)\, d\theta} \tag{30}$$

where $\hat{p}(\theta|\Lambda_{-m}, \ldots \Lambda_0)$ is calculated by choosing a sequence of sets of "a priori observations," denoted by $\Lambda_{-m}, \ldots \Lambda_0$,* and applying Eq. (20), and where $r(\theta)$ is a non-negative, integrable, but otherwise arbitrary function of θ.**

Conversely, if an a priori $p(\theta)$ of the form in Eq. (30) be assumed, there results for the a posteriori density

$$p(\theta|\Lambda_1, \ldots \Lambda_n) = \frac{p(\Lambda_1, \ldots \Lambda_n|\theta)\, p(\theta)}{\int p(\Lambda_1, \ldots \Lambda_n|\theta)\, p(\theta)\, d\theta}$$

$$= \frac{\hat{p}(\theta|\Lambda_{-m}, \ldots \Lambda_0, \Lambda_1, \ldots \Lambda_n)\, r(\theta)}{\int \hat{p}(\theta|\Lambda_{-m}, \ldots \Lambda_0, \Lambda_1, \ldots \Lambda_n)\, r(\theta)\, d\theta} \tag{31}$$

where use has been made of Eqs. (20) and (30) and of the assumption that the Λ_i's are conditionally independent given θ. If a sufficient statistic for θ of fixed dimension exists, the same analysis used in deriving Theorem III shows that both Eqs. (30) and (31) are of the same form, and hence that $p(\theta)$ is a reproducing-type a priori density.

---

*The "a priori observations" are utilized to represent the available a priori information. In a typical application the sets $\Lambda_{-m}, \ldots \Lambda_0$ are sets which are thought a priori to be typical sets of observations, with the total number of observations in these sets a measure of the confidence placed in the a priori information (see Section F). Actually, of course, only the sufficient statistics for the a priori observations need be chosen; it is even possible to use sufficient statistics that do not correspond to physically realizable sets of observations (for example, a component of the sufficient statistics corresponding to the number of observations might not be an integer) if the form of the probability density $\hat{p}(\theta|\Lambda_{-m}, \ldots \Lambda_0)$ is unchanged. If the observations are not physically realizable, the notation of Eq. (30) may be slightly misleading; it is kept for the aid in visualizing methods of generating reproducing densities which it provides.

**Rather than stating that $r(\theta)$ itself is integrable, it would be more accurate to state that the integral in the denominator of Eq. (30) exists. It will also be assumed that similar integrals, such as those in the denominator of Eq. (31), exist.

The following theorem has now been proved:

Theorem IV: Assume that the sets of observations Λ_i and Λ_j, $i \neq j$, are conditionally independent given θ. Then a reproducing-type a priori density $p(\theta)$ exists if and only if a sufficient statistic for θ of fixed dimension exists. Any reproducing-type density that exists is of the form given in Eq. (30).

Theorem IV is the fundamental theorem in the analysis of reproducing-type densities in the case where the conditional independence assumption is satisfied. It indicates that the learning process can satisfy the definition of feasibility utilized in this report (see Chapter III, Section D) if and only if a sufficient statistic of a simple form exists. It also gives a method for generating any reproducing-type densities that do exist. All those that exist can be generated by taking a function of θ of the form of the likelihood, $p(\Lambda_{-m}, \ldots \Lambda_0|\theta)$ of possible sets of observations, multiplying by an arbitrary non-negative function of θ, and then normalizing. In deriving Eq. (30), this normalization was done in two steps, first normalizing $p(\Lambda_{-m}, \ldots \Lambda_0|\theta)$ to obtain $\hat{p}(\theta|\Lambda_{-m}, \ldots \Lambda_0)$, then multiplying by $r(\theta)$ and renormalizing. A one-step normalization will suffice, as putting the definition of $\hat{p}(\theta|\Lambda_{-m}, \ldots \Lambda_0)$ [Eq. (20)] into Eq. (30) gives

$$p(\theta) = \frac{p(\Lambda_{-m}, \ldots \Lambda_0|\theta)\, r(\theta)}{\int p(\Lambda_{-m}, \ldots \Lambda_0|\theta)\, r(\theta)\, d\theta} \tag{30a}$$

Similarly, Eq. (31) may be rewritten as

$$p(\theta|\Lambda_1, \ldots \Lambda_n) = \frac{p(\Lambda_{-m}, \ldots \Lambda_0, \Lambda_1, \ldots \Lambda_n)\, r(\theta)}{\int p(\Lambda_{-m}, \ldots \Lambda_0, \Lambda_1, \ldots \Lambda_n)\, r(\theta)\, d\theta} \tag{31a}$$

The existence of a sufficient statistic of fixed dimension is more important than the use of a reproducing-type a priori density as a criterion for determining the feasibility of the learning process. In fact, the same arguments used to establish Theorem IV can be used to establish the following:

Theorem V: Assume that the sets of learning observations Λ_i and Λ_j, $i \neq j$, are conditionally independent given θ. Then, regardless of the a priori density $p(\theta)$, the density $p(\theta|\Lambda_1)$ reproduces itself with respect to the likelihood $p(\Lambda_2|\theta)$ if and only if a sufficient statistic for θ of fixed dimension exists.

Thus, if there is no objection to one reprogramming of the learning system after the first set of learning observations, it is merely necessary that there exist a sufficient statistic of fixed dimension. The form of the learning system will remain fixed after this one change, regardless of what a priori $p(\theta)$ is used. It may not always be obvious that the form is constant, but it will be possible algebraically to manipulate the densities into the form in Eq. (19). Since $\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n)$ remains of constant form, the whole density in Eq. (19) remains of constant form.

Another result similar to that in Theorem V should be pointed out. Regardless of what a priori density $p(\theta)$ is used, it is always possible to write the density in the form of Eq. (30a), i.e., as a reproducing density. To do this, it is merely necessary to pick an arbitrary sequence $\Lambda_{-m}, \ldots \Lambda_0$ of sets of " a priori observations" and multiply both numerator and denominator of the a priori density by $p(\Lambda_{-m}, \ldots \Lambda_0|\theta)$. Rewriting the density in this manner appears physically meaningless, however. Also, in view of Theorem V, little appears to be gained by such an approach. Although this possibility should be noted, it will normally be neglected in this report. Unless otherwise stated, it is assumed that the denominator of $r(\theta)$ contains no terms of the form of the likelihood function.

E. CONVERGENCE RATES WITH VARIOUS A PRIORI DENSITIES

In view of Theorem V, it appears that the use of nonreproducing a priori distributions will often give little if any increase in the complexity of the learning system. If the rate at which the a posteriori density approached a delta function were greater with a nonreproducing a priori density, the latter type of a priori density might be preferred despite some small increase in complexity. It is

easy to prove that no appreciable increase in rate of convergence can be obtained by choosing a different a priori probability density, however; the proof follows.

Consider two a priori densities $p_o(\theta)$ and $p_1(\theta)$ and the corresponding a posteriori densities $p_o(\theta|\Lambda_1, \ldots \Lambda_n)$ and $p_1(\theta|\Lambda_1, \ldots \Lambda_n)$. If $p_o(\theta)$ and $p_1(\theta)$ are approximately the same width, then $p_o(\theta|\Lambda_1, \ldots \Lambda_n)$ and $p_1(\theta|\Lambda_1, \ldots \Lambda_n)$ are approximately the same width. To show this, it is assumed that $p_o(\theta)$ and $p_1(\theta)$ both have the same mode* θ_o and that for some other point θ_1

$$\frac{p_o(\theta_o)}{p_o(\theta_1)} = \frac{p_1(\theta_o)}{p_1(\theta_1)} \tag{32}$$

(where θ_1 might be a common 3-db point for the a priori densities). Then, from Eqs. (19) and (32)

$$\frac{p_o(\theta_o|\Lambda_1, \ldots \Lambda_n)}{p_o(\theta_1|\Lambda_1, \ldots \Lambda_n)} = \frac{\hat{p}(\theta_o|\Lambda_1, \ldots \Lambda_n)}{\hat{p}(\theta_1|\Lambda_1, \ldots \Lambda_n)} \cdot \frac{p_o(\theta_o)}{p_o(\theta_1)}$$

$$= \frac{\hat{p}(\theta_o|\Lambda_1, \ldots \Lambda_n)}{\hat{p}(\theta_1|\Lambda_1, \ldots \Lambda_n)} \cdot \frac{p_1(\theta_o)}{p_1(\theta_1)} = \frac{p_1(\theta_o|\Lambda_1, \ldots \Lambda_n)}{p_1(\theta_1|\Lambda_1, \ldots \Lambda_n)} \tag{33}$$

Hence, the two a posteriori densities narrow down equally fast as more observations are taken.

F. GENERALIZATION OF THE THEORY TO INCLUDE DEPENDENT LEARNING OBSERVATIONS

The results may now be generalized to apply to the case where the learning observations are not necessarily independent given θ. The procedure will be first to give a simple example of finding a reproducing density without the assumption of conditional independence, then to use

*The mode of the density is the value of θ for which the density takes its maximum value.

this example to deduce the changes necessary in the theory in order to cover the general case.

A binary Markov process is a simple example of a case where the observations are not conditionally independent given the parameters characterizing the process. If it is assumed there are two possible states, 1 and 0, and if P_{ij} is assumed to be the probability of a transition from state i to state j, the probability of observing a 1 or 0 at a given time is not a function of the P_{ij}'s alone. It also depends on the previous digits observed. Hence, the theory thus far developed is not directly applicable.

Reproducing densities for the P_{ij}'s can easily be found, however.* If each λ_i consists of a single observation and the sequence $\{\lambda_1, \ldots \lambda_n\}$ contains a total of n_1 ones, of which r_{11} are followed by ones, and n_0 zeroes, r_{00} followed by zeros, there results,

$$P(\lambda_1, \ldots \lambda_n | P_{00}, P_{11}) = P(\lambda_1) P_{00}^{r_{00}} (1-P_{00})^{n_0 - r_{00}} P_{11}^{r_{11}} (1-P_{11})^{n_1 - r_{11}} \quad (34)$$

where use has been made of the fact that $P_{i0} + P_{i1} = 1$.

A reproducing-type density can be found for this case in the same manner as before, picking the "a priori observations" $\{\lambda_{-m}, \ldots \lambda_0\}$ consisting of n_1' ones, r_{11}' followed by ones, and n_0' zeros, r_{00}' followed by zeros, and setting

*In this case the learning observations are discrete random variables, while the theory has been developed assuming the observations were continuous random variables. There is no difficulty in extending the theory to allow observations which are discrete random variables, however. The only change necessary is replacing probability-density functions by probability-mass functions in the equations; this may be verified by replacing Eq. (2) by the form of Bayes' rule applicable here, and developing the theory in an identical manner.

$$p_{\Lambda_0}(P_{00}, P_{11}) = \frac{P(\Lambda_{-m}) P_{00}^{r'_{00}} (1-P_{00})^{n'_0 - r'_{00}} P_{11}^{r'_{11}} (1-P_{11})^{n'_1 - r'_{11}}}{\int_0^1 \int_0^1 P(\Lambda_{-m}) P_{00}^{r'_{00}} (1-P_{00})^{n'_0 - r'_{00}} P_{11}^{r'_{11}} (1-P_{11})^{n'_1 - r'_{11}} \, dP_{00} \, dP_{11}} \tag{35}$$

The parameter Λ_0 has been included as an index for the density in Eq. (35) since the computation to be performed after observing Λ_1 depends on Λ_0. For example, if Λ_1 is a one and Λ_0 a zero, then

$$p_{\Lambda_0}(P_{00}, P_{11} | \Lambda_1) = \frac{P(\Lambda_{-m}) P_{00}^{r'_{00}} (1-P_{00})^{n'_0 - r'_{00}+1} P_{11}^{r'_{11}} (1-P_{11})^{n'_1 - r'_{11}}}{\int_0^1 \int_0^1 P(\Lambda_{-m}) P_{00}^{r'_{00}} (1-P_{00})^{n'_0 - r'_{00}+1} P_{11}^{r'_{11}} (1-P_{11})^{n'_1 - r'_{11}} \, dP_{00} dP_{11}} \tag{36}$$

while if Λ_1 is a one and Λ_0 also a one

$$p_{\Lambda_0}(P_{00}, P_{11} | \Lambda_1) = \frac{P(\Lambda_{-m}) P_{00}^{r'_{00}} (1-P_{00})^{n'_0 - r'_{00}} P_{11}^{r'_{11}+1} (1-P_{11})^{n'_1 - r'_{11}}}{\int_0^1 \int_0^1 P(\Lambda_{-m}) P_{00}^{r'_{00}} (1-P_{00})^{n'_0 - r'_{00}} P_{11}^{r'_{11}+1} (1-P_{11})^{n'_1 - r'_{11}} \, dP_{00} dP_{11}} \tag{37}$$

The two expressions, Eqs. (36) and (37) differ in the exponent which is increased to allow for the additional observation. The computations after observing Λ_2 will differ similarly according to whether Λ_1 is a one or a zero. However, the densities will always be of the form in Eq. (35), so the density reproduces.

In the case of more general types of dependence, a similar procedure can be used; although the computation to be performed may depend on more than the immediately-preceding digit. Such a situation is treated by

introducing a parameter α_i, which indicates the state of the system after the i^{th} observation. In the most general case α_i may reflect the complete past history of the system. Using this parameter to index the densities,

$$p_{\alpha_0}(\theta|\Lambda_1, \ldots \Lambda_n) = \frac{p_{\alpha_0}(\Lambda_1, \ldots \Lambda_n|\theta)p_{\alpha_0}(\theta)}{\int p_{\alpha_0}(\Lambda_1, \ldots \Lambda_n|\theta)\, p_{\alpha_0}(\theta)\, d\theta} \tag{38}$$

If the original density is of the form

$$p_{\alpha_0}(\theta) = \frac{p_{\alpha_{-m}}(\Lambda_{-m}, \ldots \Lambda_0|\theta)\, r(\theta)}{\int p_{\alpha_{-m}}(\Lambda_{-m}, \ldots \Lambda_0|\theta)\, r(\theta)d\theta} \tag{39}$$

it is found that:

$$p_{\alpha_0}(\theta|\Lambda_1, \ldots \Lambda_n) = \frac{p_{\alpha_0}(\Lambda_1, \ldots \Lambda_n|\theta)p_{\alpha_{-m}}(\Lambda_{-m}, \ldots \Lambda_0|\theta)\, r(\theta)}{\int p_{\alpha_0}(\Lambda_1, \ldots \Lambda_n|\theta)p_{\alpha_{-m}}(\Lambda_{-m}, \ldots \Lambda_0|\theta)r(\theta)d\theta} \tag{40}$$

But since α_0 reflects the entire past history of the system, it is possible to write

$$p_{\alpha_0}(\Lambda_1, \ldots \Lambda_n|\theta) = p_{\alpha_{-m}}(\Lambda_1, \ldots \Lambda_n|\theta, \Lambda_{-m}, \ldots \Lambda_0) \tag{41}$$

By putting this expression in Eq. (40), there results

$$p_{\alpha_0}(\theta|\Lambda_1, \ldots \Lambda_n) = \frac{p_{\alpha_{-m}}(\Lambda_{-m}, \ldots \Lambda_0, \Lambda_1, \ldots \Lambda_n|\theta)\, r(\theta)}{\int p_{\alpha_{-m}}(\Lambda_{-m}, \ldots \Lambda_0, \Lambda_1, \ldots \Lambda_n|\theta)r(\theta)d\theta} \tag{42}$$

The same type of analysis as used in the case where the observations were independent given θ shows that Eqs. (39) and (42) are of the same form if and only if a sufficient statistic of fixed dimension exists.

The proof has now been completed for:

Theorem VI: A reproducing a priori density $p_{\alpha_0}(\theta)$ exists if and only if a sufficient statistic for θ of fixed dimension exists. Any reproducing density that exists is of the form shown by Eq. (39).

Even though the densities reproduce, the process may not be feasible if the α_i's can take on very many different values. There appears to be nothing in the theory that requires the number of different possible α_i's to be finite, or even countable, in order to have the densities reproduce. Such questions are largely academic, however, as different values of the α_i's normally mean different computations to determine the new density on θ (as in the binary Markov example) with corresponding changes in the form of the learning system.

It is possible to make a statement similar to Theorem V in this case also. The a posteriori densities eventually become reproducing if and only if a sufficient statistic of fixed dimension exists, no matter what a priori density is used. The densities may not begin reproducing before the system goes through all its possible states, or distinct α_i's, however.

G. DISCUSSION OF RESULTS

Solutions are now available for the second and third problems posed at the end of Chapter III: finding conditions that insure that reproducing-type densities exist, and finding methods for generating any reproducing-type densities that do exist. It has been shown that the existence of a sufficient statistic of fixed dimension guarantees the existence of reproducing densities, and that any reproducing densities that exist can be generated by normalizing a non-negative function of θ having a factor of the form of the likelihood of a possible set of observations. The existence of a suitable sufficient statistic is more important than the use of reproducing distributions in insuring the feasibility of the learning process, as the sequence of a posteriori distributions eventually becomes reproducing if such a statistic exists, regardless of the a priori distribution. No appreciable increase

in rate of convergence of the a posteriori densities to a delta function can be obtained by the use of a non-reproducing a priori density, however.

The results apply either with the learning observations conditionally independent given θ, or without this independence. Without the independence assumption, however, the form of the learning system may depend on the state of the system determined by previous observations. If many such states are possible, the learning procedure may be impractical even when reproducing-type distributions are used.

The class of reproducing-type a priori densities of the form in Eq. (30) or (39) is large enough to give considerable freedom in choosing a priori densities. The a priori observations (or the sufficient statistics describing these observations) can be chosen almost arbitrarily. As the examples in the next chapter show, this allows considerable freedom in choosing the "experimental portion" of the a priori density.

The function $r(\theta)$ can also be used to incorporate a wide variety of forms of a priori information. Although any non-negative function of θ can be used for $r(\theta)$ (assuming the integrability requirements over θ are met), most of these forms are physically meaningless. In the next chapter are given a few examples of forms that $r(\theta)$ may take. One of the more interesting forms for $r(\theta)$ is a constant. When $r(\theta)$ is constant, the a priori density in Eq. (30) or (39) is identical to the a posteriori density that would have been obtained after actually observing the "a priori observations," if a uniform a priori density had been assumed.*

The a priori knowledge reflected by densities of the forms in Eqs. (30) or (39) may be considered to be of two forms: one form equivalent to knowledge that could have been obtained from observations and the other form representing knowledge that could not have been obtained in this manner. Thus, all the knowledge about θ incorporated in the

*This argument breaks down if θ is defined over a set of infinite Lebesgue measure, since uniform densities over sets of infinite measure have no meaning in the conventional theory of probability. Such densities do have meaning in the theory developed by Renyi [Ref. 21], however.

"experimental portion" of the a priori density, $\hat{p}(\theta|\Lambda_{-m}, \ldots \Lambda_0)$, could have been obtained from actual observations; this is not necessarily true of the knowledge incorporated in $r(\theta)$, however.

A simple measure of confidence in the a priori knowledge contained in $\hat{p}(\theta|\Lambda_{-m}, \ldots \Lambda_0)$ is available. The confidence placed in the portion of the a priori knowledge reflected in the "experimental portion" of the a priori density is considered proportional to the size of the set of observations necessary to generate this portion of the density. In each case that has been examined (see Chapter VI), this experimental portion of the density approaches a uniform density as the size of the set of observations approaches zero, and approaches a delta function as the size increases without limit. These are the limits that would be expected as the amount of a priori knowledge approached zero or approached complete knowledge of θ, respectively.

H. USE OF BAYES' RULE COMPUTER

By applying the factorization theorem for sufficient statistics, Eq. (31) can be rewritten as:

$$p(\theta|\Lambda_1, \ldots \Lambda_n) = \frac{f(t_1^{(-m,n)}, \ldots t_s^{(-m,n)}, \theta)\ r(\theta)}{\int f(t_1^{(-m,n)}, \ldots t_s^{(-m,n)}, \theta)\ r(\theta)\ d\theta} \qquad (43)$$

where the $t_i^{(-m,n)}$ are the components of the sufficient statistic for the combined a priori and a posteriori observations. Since $r(\theta)$ is a fixed function of θ, the density in Eq. (43) is a fixed function of θ and the parameters $t_1^{(-m,n)}, \ldots t_s^{(-m,n)}$. Combining this with the previous results gives the schematic diagram drawn in Fig. 4 for the Bayes' rule computer in Fig. 1. If reproducing a priori densities are not used, the form of the computer may change initially, but will eventually become that in Fig. 4.

By incorporating the form of the Bayes' rule computer shown in Fig. 4 in the model of Fig. 1, a more detailed model for the learning process is obtained with conditionally independent observations. The chief

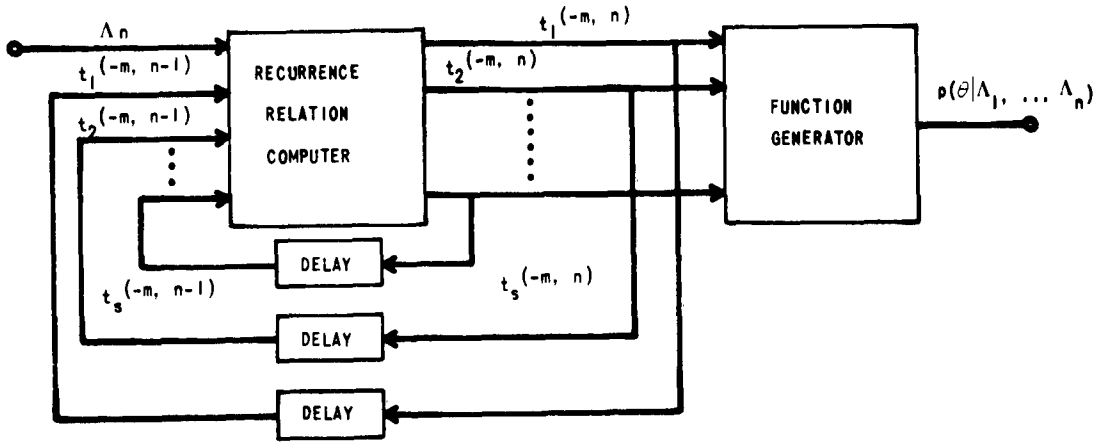

FIG. 4. BAYES' RULE COMPUTER WITH REPRODUCING A PRIORI DENSITY.

difference in the model if it were designed for the case without conditional independence would be that the form of the Bayes' rule computer might depend on the value of α_n. If it be assumed that α_n may take on r possible values, the learning process can be illustrated by the model shown in Fig. 5. The computer selector in this model computes the value of α_n and feeds Λ_n into the appropriate Bayes' rule computer. If the learning observations are conditionally independent given θ, the model in Fig. 5 reduces to that in Fig. 1, since in this case α_n may be considered to be constant.

Rather than using different Bayes' rule computers for different states of the learning system, it may well be more practical to use one computer with a variable program. If this approach is used, the computer selector in Fig. 5 may be considered to be a computation program selector. The same model applies with some minor relabeling.

In all the theory that has been developed, it has been assumed that the equations deal with probability densities only, for the sake of convenience. Any of the densities can be replaced by probability mass functions if discrete rather than continuous random variables are

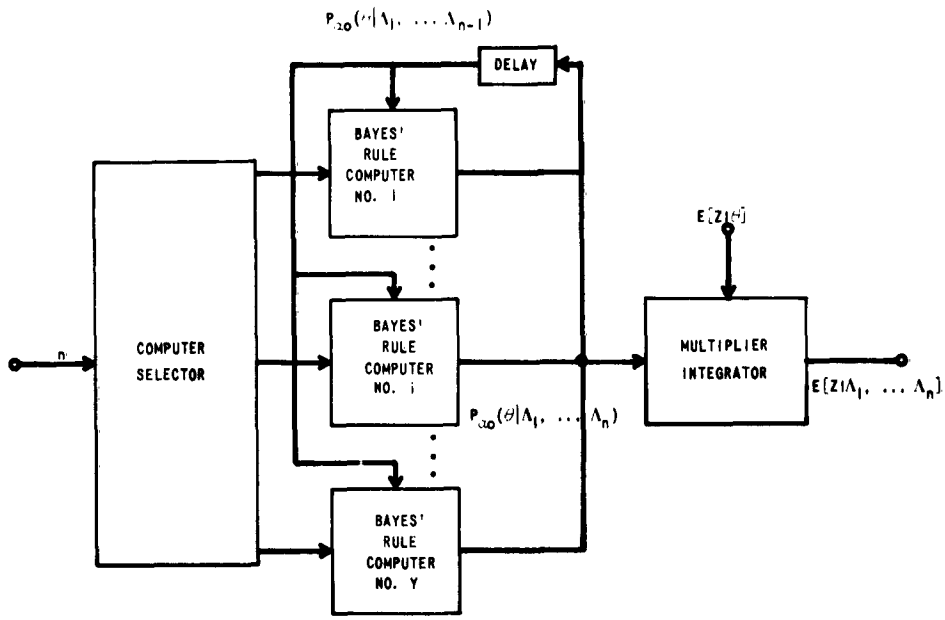

FIG. 5. GENERAL MODEL FOR LEARNING PROCESS.

encountered.* Some of the alternate equations have actually been utilized in the example introducing the methods of generalizing to the case where the learning observations are dependent.

The next chapter is devoted to examples of reproducing-type distributions. These examples should clarify some of the theory developed in the investigation.

*The term "reproducing-type distributions" is used in the title of this report as being more general than "reproducing-type densities." Probability mass functions may reproduce also.

VI. EXAMPLES OF REPRODUCING-TYPE DISTRIBUTIONS

In this chapter are given a number of examples of probability distributions that are reproducing. The two criteria that have been utilized in choosing the examples are the engineering utility of the probability distributions involved and the possibility of illustrating different properties of the distributions.

Two different classes of reproducing distributions are considered. For the first class, called simple reproducing distributions, $r(\theta)$ is a constant and hence $p(\theta)$ equals $\hat{p}(\theta|\Lambda_{-m}, \ldots \Lambda_0)$. For the second class, called composite reproducing distributions, $r(\theta)$ is not constant. Hence, a composite reproducing distribution is the product of a simple reproducing distribution and another function of θ.

A. A SAMPLE COMPUTATION: THE BINOMIAL DISTRIBUTION

The binomial distribution is probably the most common discrete probability distribution in engineering applications. It might be termed, in everyday engineering language, the "go--no go" distribution. This distribution can describe the probability that a switch is open or closed; or the probability that a signal corresponds to a one or to a zero; or a myriad of other cases where only two events are considered to be possible. If the probability P characterizing this distribution is unknown, the learning procedure developed in this paper is applicable.

It is assumed for the sake of definiteness that the two possible events are the reception of a one and of a zero. If P were known, it would be the probability of a one. Each Λ_i is assumed to be the observation of a single digit.

To find a simple reproducing density, a specific a priori sequence $\Lambda_{-n_o+1}, \ldots \Lambda_0$ consisting of r_o ones and $n_o - r_o$ zeros is assumed. Making use of Theorem II, Chapter V, and the basic definition given by Eq. (20), but replacing the symbols $p(\Lambda_i, \ldots \Lambda_j|\theta)$ for the likeliho d functions by the discrete random variable analogs $P(\Lambda_i, \ldots \Lambda_j|\theta)$ (since the binomial distribution is a discrete rather than a continuous distribution), $p(P)$ is chosen to be

$$p(P) = p(P|\Lambda_{-n_o+1}, \cdots \Lambda_0) = \frac{P(\Lambda_{-n_o+1}, \cdots \Lambda_0|P)}{\int P(\Lambda_{-n_o+1}, \cdots \Lambda_0|P)\, dP}$$

$$= \begin{cases} \dfrac{P^{r_o}(1-P)^{n_o-r_o}}{\int_0^1 P^{r_o}(1-P)^{n_o-r_o}\, dP} = \dfrac{\Gamma(n_o+2)}{\Gamma(r_o+1)\Gamma(n_o-r_o+1)} P^{r_o}(1-P)^{n_o-r_o}, & 0 \le P \le 1, \\ 0, & \text{otherwise.} \end{cases} \quad (44)$$

The density given by Eq. (44) may be recognized as a beta density function. This fact can be used to check the normalizing constant obtained. Alternatively, the normalizing constant can be obtained by finding a standard probability density function that depends on its argument in the same way that the function in Eq. (44) depends on P--relying on the fact that standard density functions are normalized to integrate to one. In any event, determining whether the density is a standard form is useful, since, if such is the case, the important properties of the density may have been tabulated.

In the equations for the a posteriori density when a reproducing a priori density is used, there is no distinction between effects of a priori and a posteriori observations. Hence, the a posteriori density after observing a sequence consisting of r_1 ones and n_1-r_1 zeros is

$$p(P|\Lambda_1, \cdots \Lambda_{n_1}) = \hat{p}(P|\Lambda_{-n_o+1}, \cdots \Lambda_0, \Lambda_1, \cdots \Lambda_{n_1})$$

$$= \begin{cases} \dfrac{\Gamma(n_o + n_1 + 2)}{\Gamma(r_o+r_1+1)\,\Gamma(n_o+n_1-r_o-r_1+1)} P^{r_o+r_1}(1-P)^{n_o+n_1-r_o-r_1}, & 0 \le P \le 1 \\ 0, & \text{otherwise} \end{cases}$$

$$= \begin{cases} \dfrac{\Gamma(n+2)}{\Gamma(r+1)\,\Gamma(n-r+1)}\, P^r\,(1-P)^{n-r}, & 0 \le P \le 1 \\ 0, & \text{otherwise.} \end{cases} \tag{45}$$

where

$$n \triangleq n_o + n_1, \tag{46a}$$

$$r \triangleq r_o + r_1. \tag{46b}$$

The mean and variance of Eq. (45) are given by

$$E[P|\Lambda_1, \ldots \Lambda_n] = \frac{r+1}{n+2} \tag{47a}$$

$$\text{Var}\,[P|\Lambda_1, \ldots \Lambda_n] = \frac{(r+1)(n-r+1)}{(n+2)^2\,(n+3)} \tag{47b}$$

As the total number of a priori and a posteriori observations approaches zero, the above values approach

$$E[P|\Lambda_1, \ldots \Lambda_n] \to \frac{1}{2} \tag{48a}$$

$$\text{Var}\,[P|\Lambda_1, \ldots \Lambda_n] \to \frac{1}{12} \tag{48b}$$

These are the values of the mean and variance of a uniform density over the interval from zero to one. Conversely, as the total number of a priori and a posteriori observations becomes very large,

$$E[P|\Lambda_1, \ldots \Lambda_n] \to \lim_{r,n\to\infty} \frac{r}{n} \triangleq P_o \tag{49a}$$

$$\text{Var}\,[P|\Lambda_1, \ldots \Lambda_n] \to 0 \tag{49b}$$

These are the values of the mean and variance of a delta function density at P equals P_o. Moreover, for any finite number of a priori observations, the limiting ratio in Eq. (49a) will be the limiting ratio of the

values for the a posteriori observations, which, according to the strong law of large numbers, is, with probability one, the true value of P.

In this case it is easy to show that the limiting forms of the density, for small and large numbers of observations, are a uniform density over the interval from zero to one and a delta function at the true value of P. The results are left in the form of Eqs. (48) and (49) for easy comparison with other reproducing-type densities obtained, however.

Sufficient statistics for the sequences of observations arise naturally from the analysis. The pairs of numbers (n_o, r_o), (n_1, r_1) and (n, r) are sufficient for the a priori, a posteriori and total sequences respectively.

B. SOME SIMPLE REPRODUCING-TYPE DISTRIBUTIONS

In this section, ten typical examples of simple reproducing-type distributions are analyzed. The distributions treated, the unknown parameters, and the form of the learning observations are listed in Table 1. Table 2 gives the likelihood of the learning observations and the simple reproducing-type densities.

1. Probability Distributions Considered

Four discrete distributions are treated: the binomial, the multinomial, the binary Markov, and the Poisson. In each case parameters characterizing the probability mass function are unknown. Six examples of continuous distributions with some of the parameters characterizing the probability density functions unknown are also treated. These include three examples of Gaussian densities, one multidimensional with unknown mean vector, one multidimensional with unknown covariance matrix, and one one-dimensional with a complex mean and both magnitude and phase of the mean unknown.* The three other cases are the Rayleigh, the

*In the appendix the case of a multidimensional Gaussian density with both means and covariances unknown is also treated. The simple reproducing density in this case is the composite Wishart-Gaussian density used by Keehn (see Chapter III, Section C).

TABLE 1. PROBABILITY DISTRIBUTIONS CONSIDERED.

No.	Prob. Dist.	Unknown, θ	Type of Dist.	Learning Observations
1	Binomial	P	Discrete	r ones, n-r $\triangleq$ s zeros
2	Multinomial	$P_1, P_2, \ldots, P_{m-1}$	Discrete	r_1 ones, r_2 twos, ..., r_{m-1} (m-1)'s $n-r_1 \ldots -r_{m-1} \triangleq r_m$ zeros
3	Binary Markov	P_{00} , P_{11}	Discrete	sequence with n_1 ones, r_{11} followed by one; n_0 zeroes, r_{00} followed by zero
4	Poisson	α	Discrete	n events in time τ
5	Gaussian	M	Continuous	n samples each sample d-dimensional, a vector in R^d
6	Gaussian	K^{-1}	Continuous	n samples each sample d-dimensional, a vector in R^d
7	Complex Gaussian	a, φ	Continuous	n samples each sample complex number
8	Rayleigh	$\rho \triangleq \frac{1}{\sigma^2}$	Continuous	n samples each sample non-negative number
9	Exponential	λ	Continuous	n samples each sample non-negative number
10	Zero-mean rectangular	W	Continuous	n samples each sample real number, $\| \| \leq W$

TABLE 2. SIMPLE REPRODUCING DENSITIES.

No.	Likelihood of Observations	Simple Reproducing Density
1	$P^r(1-P)^s$	$\frac{\Gamma(n+2)}{\Gamma(r+1)\Gamma(s+1)} P^r(1-P)^s$, $0 \le P \le 1$ (beta) 0, otherwise
2	$P_1^{r_1} P_2^{r_2} \cdots P_m^{r_m}$, $P_m \triangleq 1 - P_1 - \cdots - P_{m-1}$	$\frac{\Gamma(n+m)}{\Gamma(r_1+1),..\Gamma(r_m+1)} P_1^{r_1} P_2^{r_2} \cdots P_m^{r_m}$, $P_1, \ldots, P_{m-1}$ in simplex S $S = \{P_1, \ldots, P_{m-1}\}: P_i \ge 0 \sum P_i \le 1\}$ 0, otherwise (Dirichlet)
3	P(1st digit) $P_{11}^{r_{11}}(1-P_{11})^{r_{10}} P_{00}^{r_{00}} (1-P_{00})^{r_{01}}$ $r_{10} \triangleq n_1 - r_{11}$, $r_{01} \triangleq n_0 - r_{00}$	$\frac{\Gamma(n_1+2)}{\Gamma(r_{11}+1)\Gamma(r_{10}+1)} P_{11}^{r_{11}}(1-P_{11})^{r_{10}} \frac{\Gamma(n_0+2)}{\Gamma(r_{00}+1)\Gamma(r_{01}+1)} P_{00}^{r_{00}}(1-P_{00})^{r_{01}}$ $0 \le P_{00}, P_{11} \le 1$ (double beta)* 0, otherwise
4	$\frac{(\alpha\tau)^n}{n!} e^{-\alpha\tau}$	$\frac{1}{n!}(\alpha\tau)^n e^{-\alpha\tau}$, $\tau \ge 0$ (gamma) 0, otherwise
5	$\frac{1}{(2\pi)^{d/2}\lvert K\rvert^{n/2}} \exp[-\frac{1}{2}\Sigma(X_i-M)_t K^{-1}(X_i-M)]$	$\frac{1}{(2\pi)^{d/2}\lvert K_n\rvert^{1/2}} \exp[-\frac{1}{2}(M-\bar{X}_n)_t K_n^{-1}(M-\bar{X}_n)]$, $\bar{X}_n \triangleq \frac{1}{n}\Sigma X_i$ (Gaussian) $K_n \triangleq \frac{1}{n} K$
6	$\lvert(2\pi)^d K\rvert^{-n/2} \exp(-\frac{1}{2}\operatorname{tr} V_n K^{-1})$ $V_n \triangleq \Sigma (X_i - M)(X_i - M)_t$	$\frac{\lvert V_n\rvert^{\frac{n+d+1}{2}} \lvert K\rvert^{-n/2} \exp(-\frac{1}{2}\operatorname{tr} V_n K^{-1})}{2^{\frac{1}{2}nd}\pi^{\frac{1}{4}d(d-1)} \prod_{i=0}^{d-1} \Gamma(\frac{n+d+1-i}{2})}$, K^{-1} positive definite and symmetric 0, otherwise (Wishart)
7	$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp[-\frac{1}{2\sigma^2}\Sigma\lvert X_i - \alpha e^{j\phi}\rvert^2]$ $= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp[-\frac{1}{2\sigma^2}\Sigma(\lvert X_i\rvert^2 + 2\alpha\lvert X_i\rvert\cos(\phi-\alpha_i)+\alpha^2)]$ $\alpha_i \triangleq \tan^{-1}\frac{\operatorname{Im}(X_i)}{\operatorname{Re}(X_i)}$	$\frac{1}{2^{\frac{1}{2}}\pi^{\frac{1}{4}}\sigma_n} I_0\left[\frac{\lvert\bar{X}_n\rvert^2}{4\sigma_n^2}\right] \exp[-\frac{1}{2\sigma_n^2}(\alpha^2 - 2\alpha\lvert\bar{X}_n\rvert\cos(\phi-\eta)+\lvert\bar{X}_n\rvert^2)]$ $\alpha \ge 0$, $-\pi \le \phi \le \pi$ 0, otherwise $\eta_n \triangleq \tan^{-1}\frac{\Sigma \operatorname{Im}(X_i)}{\Sigma \operatorname{Re}(X_i)}$ $\lvert\bar{X}_n\rvert \triangleq \frac{1}{n}\lvert\Sigma X_i\rvert$ $\sigma_n^2 \triangleq \frac{1}{n}\sigma^2$
8	$\alpha^n(\Pi X_i)\exp(-\frac{1}{2}\Sigma X_i^2)$	$\frac{K_n}{n!}(k_n\alpha)^n \exp(-k_n\alpha)$, $\alpha \ge 0$, $K_n \triangleq \frac{1}{2}\Sigma X_i^2$ (gamma) 0, otherwise
9	$\lambda^n \exp(-\lambda \Sigma X_i)$	$\frac{C_n}{n!}(C_n\lambda)^n \exp(-C_n\lambda)$, $\lambda \ge 0$, $C_n \triangleq \Sigma X_i$ (gamma) 0, otherwise
10	$\left(\frac{1}{2W}\right)^n u(W-M_n)$, $u(X) \triangleq \begin{cases}1, & X \ge 0\\ 0, & X < 0\end{cases}$ $M_n \triangleq \max\lvert X_i\rvert$	$\frac{n-1}{W_n}\left(\frac{W_n}{W}\right)^n u(W-W_n)$ $n > 1$ undefined, $n \le 1$

*This assumes the first digit is chosen independently of P_{00} and P_{11}.

exponential, and the zero-mean rectangular distributions with parameters characterizing these distributions unknown.

Each of the ten distributions considered has important engineering applications. The binomial, multinomial, and binary Markov distributions are important in such fields as coding, hypothesis testing, and pattern recognition. Typical applications of the Poisson distribution are in the study of shot noise and various waiting time and counting problems. The Gaussian densities occur so often that little comment is necessary, save for the fact that the form with a complex mean is the form that would be used when using complex numbers to indicate both magnitude and phase information in a single number. The Rayleigh density is the probability density for the envelope of a narrow-band Gaussian random process and (among other applications) is used in the study of the fading of radio signals. The exponential density is the density for the output of a square-law detector (square-law device followed by a low-pass filter), with a narrow-band Gaussian input. The final case, the rectangular density, is useful in such areas as the study of systems with unknown phases or an unknown time reference, or studies involving the location of an object confined to a specific interval.

2. Computation Methods

In computing the reproducing densities for Table 2, subscripts to indicate that the observations are "a priori observations" have been omitted. The densities may be considered as either a priori or a posteriori forms, since a priori and a posteriori observations are equivalent in their effects on the densities.

Each of the densities in Table 2 was obtained in a manner analogous to the computation for the binomial distribution given in the previous section. In two cases--the Gaussian with unknown covariances (Case 6) and the Rayleigh (Case 8)--it was found convenient to define as a new parameter the inverse of the unknown, and then to find a reproducing density for this inverse parameter. This was done purely for the sake of convenience; by writing the densities in terms of the inverse parameters ρ and K^{-1} standard forms are obtained with the normalization constants and important properties tabulated. In each of the

eight cases where standard probability densities were obtained as the reproducing densities, the common name for the density obtained is indicated in Table 2.

3. Analysis of Reproducing Densities

The first case on the list (the binomial distribution) has already been discussed in some detail. The second case, multinomial distribution with P_i's unknown, and the third case, binary Markov with P_{ii}'s unknown, are generalizations of the binomial case. It is found that the reproducing density for the multinomial distribution (which is equivalent to the (m-1)-dimensional generalization of the binomial distribution) is the (m-1)-dimensional generalization of the beta density, i.e., it is the Dirichlet density. Similarly, in the binary Markov case, by assuming that the first digit of the a priori sequence for learning the unknown P_{00} and P_{11} is chosen independently of P_{00} and P_{11}, any interaction between these two probabilities is removed, so that they can be treated as independent random variables, each distributed according to a beta density.

The three cases discussed above--binomial, multinomial and binary Markov--may be encountered in determining thresholds for likelihood ratio tests in pattern recognition. It is possible, moreover, to utilize these learning techniques to obtain the thresholds. This may result in using variable thrcholds. This possibility is discussed in more detail in the next chapter.

The binary Markov process is an example of a case where a reproducing-type density can be found without assuming that the Λ_i are conditionally independent given θ. This is the case that was utilized to introduce the method of generalizing to allow for dependent learning observations in Chapter V, Section D. It is the only example included herein in which the learning observations are not conditionally independent given θ. Other cases of this type can be treated in an analogous manner, although most of them will be more complex.

The densities obtained for the multivariate Gaussian process with unknown mean vector (Case 5) and with unknown covariance matrix (Case 6), and for the case with both mean vector and covariance matrix unknown

(which is included in the appendix), are the densities that Abramson, Braverman and Keehn have shown to be of the reproducing type as discussed in Chapter III. Similarly, the densities given for the binomial and multinomial cases are those used by Bellman and Mosimann, respectively, and a number of the densities have been used by Raiffa and Schlaifer. The only case mentioned in Chapter III for which it has been found that reproducing-type densities have been used but in which the density used is not the form in Table 2 is that discussed by Turin [Ref. 13]. The density given in Table 2 for the unknown amplitude and phase of a complex Gaussian mean is not the Rician density used by Turin, although it is similar. The difference is discussed in more detail in later sections of this chapter.

The density given in Table 2 for the complex Gaussian case (Case 7) is not as complex as it may at first seem. The density is actually simple save for the normalizing constant. This can be seen by rewriting the density in either of the forms

$$p(a,\phi) = \begin{cases} K_1 \exp\left\{-\frac{1}{2\sigma_n^2}\left[a^2 - 2a|\bar{X}_n|\cos(\phi - \delta_n)\right]\right\} & \text{or} \\ K_2 \exp\left\{-\frac{1}{2\sigma_n^2}|\bar{X}_n - a e^{i\phi}|^2\right\} & a \geq 0,\ -\pi \leq \phi < \pi \\ 0, & \text{otherwise.} \end{cases} \tag{50}$$

with K_1 and K_2 normalizing constants chosen so that either of the forms of $p(a,\phi)$ in Eq. (50) integrates to one.

The final case on the list--rectangular distribution with unknown mean--is a rather off-beat example. This density violates some of the statistical criteria for "regularity," since it is not continuous. The reproducing density obtained also has unusual properties. It is the only case encountered in this study where the density is not defined after one observation because $p(\Lambda_1|\theta)$ is not integrable. Some care

must be exercised in picking "a priori observations" also, since these must be less than W in absolute magnitude. If this condition is not fulfilled, the a priori p(W) will be zero at the true value of W and the a posteriori density cannot degenerate at the correct point. Picking observations less than W in absolute magnitude may be difficult if nothing is known about W.

4. Sufficient Statistics

Sufficient statistics for each of the various probability distributions analyzed can easily be obtained from Table 2, since the densities therein are expressed in terms of the sufficient statistics. For the binomial distribution it is found that n and r (or r and s) constitute a sufficient statistic. Similarly, for the multinomial distribution, $r_1, \ldots r_m$ are sufficient; for the binary Markov, r_{11}, n_1, r_{00} and n_0; for the Poisson, τ and n; for the multidimensional Gaussian with unknown mean vector $\mathbf{X}_n$ and n; for the multidimensional Gaussian with unknown covariance matrix $\mathbf{V}_n$ and n; for the complex Gaussian, $|\overline{X}_n|$, δ_n and n; for the Rayleigh, K_n and n; for the exponential, C_n and n; and for the rectangular density, M_n and n.

5. Representation of a Priori Knowledge

When using simple reproducing densities, such as those in Table 2, the parameters of these densities can be adjusted to reflect a priori knowledge. A priori observations are selected which, on the basis of the a priori information available, appear representative of the observations to be expected; these observations are then used to generate the reproducing density.* For example, in Case 1, if the probability of obtaining a one for a binomial distribution were expected to be about $\frac{1}{2}$, a beta density for P with r and s approximately equal would be chosen; or if the mean of a Gaussian distribution (Case 5) were expected to be near zero, a priori observations with a sample average near zero would be chosen. The degree of confidence in such a priori knowledge is reflected in the size of the total set of a priori

*Normally only the sufficient statistics for these sets of a priori observations would be selected, rather than the observations themselves.

observations, or by the magnitude of such parameters as n, r or τ in the densities in Table 2. If there is reason to be confident that the a priori knowledge is approximately correct, the parameters indicating the size of this a priori set would be large; if little confidence is reposed in the a priori knowledge, the parameters selected would be small.

In some cases, the a priori knowledge is not in the form of sufficient statistics such as those in terms of which the densities in Table 2 are defined, but the a priori knowledge is better described as consisting of approximately what the value of the unknown parameter is expected to be, plus the approximate width of the expected a priori density (or the amount of deviation from the expected value that might reasonably be allowed for). In Table 3 are listed important moments, i.e., means, variances, and covariances, for the reproducing-type densities in Table 2. These moments can be utilized to fit a priori knowledge having the forms designated.

6. Limiting Forms of Densities

The moments in Table 3 are also useful in determining limiting properties of the densities as the size of the set of a priori observations (or of the combined set of a priori and a posteriori observations) becomes very large or very small. Since the size of this set indicates the degree of confidence reposed in the a priori knowledge (or the combined a priori and a posteriori knowledge), the limiting forms would be expected to be a very narrow density approximating a delta function for a large set of observations, and a very broad density approximating a uniform density for a small set of observations. Tables 4 and 5 indicate that these are indeed the limiting forms obtained.

Table 4 indicates the limiting forms for the moments obtained with a large set of observations. In each case the means approach limiting forms that are possible values for the unknown parameters, while the variances and covariances approach zero. This indicates that a delta function is the limiting form of the density for a large set of observations.

TABLE 3. IMPORTANT MOMENTS OF REPRODUCING DENSITIES.

No.	Means	Variances	Covariances
1	$E[P] = \frac{r+1}{n+2}$	$Var[P] = \frac{(r+1)(s+1)}{(n+2)^2(n+3)}$	
2	$E[P_i] = \frac{r_i+1}{n+m}$, $i = 1, \ldots, m$	$Var[P_i] = \frac{(r_i+1)(n+m-r_i-1)}{(n+m)^2(n+m+1)}$, $i = 1, \ldots, m$	$Cov[P_i, P_j] = -\frac{(r_i+1)(r_j+1)}{(n+m)^2(n+m+1)}$, $i \neq j$; $i, j = 1, \ldots, m$
3	$E[P_{ij}] = \frac{r_{ij}+1}{n_i+2}$, $i = 0, 1$	$Var[P_{ij}] = \frac{(r_{ij}+1)(r_{ij}+1)}{(n_i+2)^2(n_i+3)}$, $i = 0, 1$; $j \neq i$	$Cov[P_{00}, P_{11}] = 0$
4	$E[\alpha] = \frac{n+1}{\tau}$	$Var[\alpha] = \frac{n+1}{\tau^2}$	
5	$E[m_i] = (\bar{X}_n)_i$, $i = 1, \ldots, d$	$Var[m_i] = (K_n)_{ii} = \frac{1}{n} k_{ii}$, $i = 1, \ldots, d$	$Cov[m_i, m_j] = (K_n)_{ij} = \frac{1}{n} k_{ij}$, $i \neq j$; $i, j = 1, \ldots, d$
6	$E[k^{ij}] = u_n^{ij}$ $a^{ij} \triangleq (A^{-1})_{ij}$ $U_n \triangleq \frac{V_n}{n+d+1}$	$Var[k^{ij}] = \frac{1}{n+d+1}[(u_n^{ij})^2 + (u_n^{ii})(u_n^{jj})]$	$Cov[k^{ij}, k^{lm}] = \frac{1}{n+d+1}[(u_n^{jl})(u_n^{jm}) + (u_n^{im})(u_n^{jl})]$
7	$E(a) = \left(\frac{2}{\pi}\right)^{1/2} \sigma_n \frac{e^{\frac{R_n^2}{2}}}{I_0[\frac{R_n^2}{2}]}$ $R_n \triangleq \frac{\lvert \bar{X}_n \rvert}{2\sigma_n}$ Others complex.	$Var[a] = 2R_n^2 \sigma_n^2 \left[1 + \frac{I_1[R_n^2]}{I_0[R_n^2]}\right] + \sigma_n^2 \left[1 + \frac{2}{\pi}\left(\frac{e^{\frac{R_n^2}{2}}}{I_0[\frac{R_n^2}{2}]}\right)^2\right]$ See appendix for approximate analysis.	
8	$E[\sigma] = \frac{n+1}{K_n} = \frac{n+1}{[\sum X_i^2]^{1/2}}$	$Var[\sigma] = \frac{n+1}{K_n^2} = \frac{n+1}{[\sum X_i^2]^2}$	
9	$E[\lambda] = \frac{n+1}{C_n} = \frac{n+1}{\sum X_i}$	$Var[\lambda] = \frac{n+1}{C_n^2} = \frac{n+1}{[\sum X_i]^2}$	
10	$E[W] = \frac{n-1}{n-2} \cdot M_n$, $n > 2$ ∞, $1 < n \leq 2$ undefined, $n \leq 1$	$Var[W] = \frac{n-1}{(n-2)^2(n-3)} M_n^2$, $n > 3$ ∞, $2 < n \leq 3$ undefined, $n \leq 2$	

TABLE 4. LARGE SAMPLE LIMITS OF MOMENTS.

No.	Parameter Limits	Means	Variances	Covariances
1	$\lim \frac{r}{n} \triangleq P_o$ $\lim n = \infty$	$E[p] \to P_o$	$Var[p] \to \frac{P_o(1-P_o)}{n} \to 0$	
2	$\lim \frac{r_i}{n} \triangleq P_{io}$ $\lim n = \infty$	$E[P_i] \to P_{io}$	$Var[P_i] \to \frac{P_{io}(1-P_{io})}{n} \to 0$	$Cov[P_i, P_j] \to -\frac{P_{io}P_{jo}}{n} \to 0$
3	$\lim \frac{r_{ii}}{n_i} \triangleq P_{iio}$ $\lim n_j = \infty$	$E[P_{ii}] \to P_{iio}$	$Var[P_{ii}] \to \frac{P_{iio}(1-P_{iio})}{n_i} \to 0$	$Cov[P_{ii}, P_{jj}] = 0$
4	$\lim \frac{n}{\tau} \triangleq \alpha_o$ $\lim \tau = \infty$	$E[\alpha] \to \alpha_o$	$Var[\alpha] \to \frac{\alpha_o}{\tau} \to 0$	
5	$\lim (\bar{X}_n)_i \triangleq m_{io}$ $\lim n = \infty$	$E[m_i] \to m_{io}$	$Var[m_i] = \frac{k_{ii}}{n} \to 0$	$Cov[m_i, m_j] = \frac{k_{ij}}{n} \to 0$
6	$\lim \frac{V_n}{n} \triangleq K_o$ $\lim n = \infty$	$E[k^{ij}] \to k_o^{ij}$	$Var[k^{ij}] \to \frac{1}{n}[(k_o^{ij})^2 + (k_o^{ii})(k_o^{jj})] \to 0$	$Cov[k^{ij}, k^{lm}] \to \frac{1}{n}[(k_o^{il})(k_o^{jm}) + (k_o^{im})(k_o^{jo})] \to 0$
7	$\lim \lvert X_n \rvert \triangleq a_o$ $\lim \delta_n \triangleq \phi_o$ $\lim n = \infty$	$E[a] \to a_o$ $E[\phi] \to \phi_o$	$Var[a] \to \sigma_n^2 = \frac{\sigma^2}{n} \to 0$ $Var[\phi] \to O(\sigma_n^2) \to 0$	$Cov[a, \phi] \to O(\sigma_n^2) \to 0$
8	$\lim \frac{\frac{1}{2}\sum X_i^2}{n} \triangleq \sigma_o^2$ $\lim n = \infty$	$E[\rho] \to \frac{1}{\sigma_o^2}$	$Var[\rho] \to \frac{1}{n}\left(\frac{1}{\sigma_o^2}\right)^2 \to 0$	
9	$\lim \frac{\sum X_i}{n} \triangleq \frac{1}{\lambda_o}$ $\lim n = \infty$	$E[\lambda] \to \lambda_o$	$Var[\lambda] \to \frac{\lambda_o^2}{n} \to 0$	
10	$\lim M_n \triangleq W_o$ $\lim n = \infty$	$E[W] \to W_o$	$Var[W] \to \left(\frac{W_o}{n}\right)^2 \to 0$	

TABLE 5. SMALL SAMPLE LIMITS OF MOMENTS.

No.	Means		Variances		Covariances		Range of $\vec{x}$
	$p(\vec{x})$	Unif. dens.*	$p(\vec{x})$	Unif. dens.*	$p(\vec{x})$	Unif. dens.*	
1	$E[P] = \frac{1}{2}$	$E[P] = \frac{1}{2}$	$Var[P] = \frac{1}{12}$	$Var[P] = \frac{1}{12}$			$P \in [0, 1]$
2	$E[P_i] = \frac{1}{m}$	$E[P_i] = \frac{1}{m}$	$Var[P_i] = \frac{m-1}{m^2(m+1)}$	$Var[P_i] = \frac{m-1}{m^2(m+1)}$	$Cov[P_i, P_i] = -\frac{1}{m^2(m-1)}$	$Cov[P_i, P_j] = -\frac{1}{m^2(m+1)}$	Simplex $P_i \geq 0$, $\sum P_i \leq 1$
3	$E[P_{ii}] = \frac{1}{2}$	$E[P_{ii}] = \frac{1}{2}$	$Var[P_{ii}] = \frac{1}{12}$	$Var[P_{ii}] = \frac{1}{12}$	$Cov[P_{ii}, P_{jj}] = 0$	$Cov[P_{ii}, P_{jj}] = 0$	Rectangle $0 \leq P_{ii} \leq 1$
4	$E[\alpha] = \alpha$	$E[\alpha] = \alpha$	$Var[\alpha] = \alpha$	$Var[\alpha] = \alpha$			$\alpha \in [0, \alpha)$
5	Undefined	Undefined	$Var[m_i] = \alpha$	$Var[m_i] = \alpha$	Undefined	Undefined	$M \in R^d$
6	$E[k^{ii}] = \alpha$ Undefined $i \neq j$	$E[k^{ii}] = \alpha$ Undefined $i \neq j$	$Var[k^{ii}] = \alpha$	$Var[k^{ii}] = \alpha$	Undefined	Undefined	$K^{-1} \in R^{d^2}$ *Pos. def.* Symmetric
7	$E[a] = \alpha$ $E[\varphi] = 0$	$E[a] = \alpha$ $E[\varphi] = 0$	$Var[a] = \alpha$ $Var[\varphi] = \frac{4\pi^2}{12}$	$Var[a] = \alpha$ $Var[\varphi] = \frac{4\pi^2}{12}$	Undefined	Undefined	$a \in [0, \alpha)$ $\varphi \in [-\pi, \pi]$
8	$E[\rho] = \alpha$	$E[\rho] = \alpha$	$Var[\rho] = \alpha$	$Var[\rho] = \alpha$			$\rho \in [0, \alpha)$
9	$E[\lambda] = \alpha$	$E[\lambda] = \alpha$	$Var[\lambda] = \alpha$	$Var[\lambda] = \alpha$			$\lambda \in [0, \alpha)$
10	$E[W] = \alpha$	$E[W] = \alpha$	$Var[W] = \alpha$	$Var[W] = \alpha$			$W \in [M, \alpha)$ $M \triangleq \lim_{n \to 0} M_n$

*Where a uniform density is not defined because of the range of $\vec{x}$ having infinite Lebesque measure, limiting values of *moments for* densities approaching uniformity are given if the limit is independent of the limiting process.

Since a priori observations and a posteriori observations are treated in identical manners, Table 4 can be used to find the limiting forms of the a posteriori densities, assuming a finite a priori set of observations and an increasingly large a posteriori set. The limiting form is in each case a delta function as before, but the location of the delta function can be stated precisely. In the Appendix it is shown that, in each case, the mean converges with probability one to the true value of the unknown parameter. Hence, the densities approach delta functions at the true values of the unknown parameters, or the learning system learns the true values exactly.

In Table 5 the limiting forms of the moments are analyzed as the size of the set of a priori observations approaches zero. In making this analysis, parameters indicating the size of the a priori set have not been confined to integer values, since the densities are defined regardless of whether these parameters are integer valued or not. The procedure used to find these limiting forms is simply to let all the parameters defining the size of the set of a priori observations approach zero, finding the limiting forms of the means, variances, and covariances whenever these limiting values are uniquely defined.

In Table 5 the limiting forms obtained for the means, variances, and covariances are compared with the means, variances, and covariances of random variables distributed according to a uniform density over the range of possible values of the unknown parameter. In some cases a uniform density is not defined over this range because the range is of infinite Lebesgue measure.* In these cases the moments tabulated are the limiting values of the moments of a sequence of random variables with probability distributions approaching a uniform distribution, if the limiting values are uniquely defined; if the limits are not uniquely defined, this is indicated in Table 5. In each case, exact agreement is found between the moments of the reproducing-type densities and the moments of uniform densities. If the moments of either are uniquely defined, the moments of the other are also uniquely defined and take the same values.

*As noted earlier uniform probability densities over sets of infinite Lebesgue measure are allowed in the theory developed by Rényi [Ref. 21], however.

Details of the computing methods for all the tables are given in the Appendix.

C. SOME COMPOSITE REPRODUCING-TYPE DISTRIBUTIONS

As indicated in the previous section, simple reproducing-type distributions contain enough adjustable parameters to give considerable freedom in choosing a priori probabilities. A number of types of a priori knowledge can be reflected in these a priori distributions, including values of the parameters that are felt to be typical and a measure of the confidence reposed in the a priori knowledge.

Even more freedom in choosing a priori distributions is available if composite reproducing-type distributions are considered. As indicated in Eq. (30), a simple reproducing-type distribution multiplied by an arbitrary (except for scale factor) non-negative function of θ is still a reproducing-type distribution. These more complex reproducing-type distributions have been defined to be composite reproducing-type distributions.

In this section no attempt is made to indicate all the possibilities of choosing composite reproducing distributions. The discussion is limited to two general classes of composite reproducing distributions.

1. Restricting the Range of θ

One class of composite reproducing distributions is useful when part of the a priori knowledge is the fact that the true value of θ is contained in some interval I. For example, it might be desired to detect a signal of unknown frequency, using a receiver of a known finite bandwidth. The probability of receiving a signal outside the frequency band accepted by the receiver would be zero. In such a case $r(\theta)$ in Eq. (30) may be taken as

$$r(\theta) = \begin{cases} 1, & \theta \in I \\ 0, & \text{otherwise} \end{cases} \qquad (51)$$

giving

$$p(\theta) = \begin{cases} \dfrac{\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n)}{\int_I \hat{p}(\theta|\Lambda_1, \ldots \Lambda_n)d\theta}, & \theta \in I \\ 0, & \text{otherwise.} \end{cases} \tag{52}$$

For example, if θ were the unknown mean m of a one-dimensional Gaussian distribution with known variance σ^2, and if it were known that $a < m < b$, an a priori density on m might be obtained by picking an a priori set $\{X_{-t+1}, \ldots X_0\}$ of learning observations (all confined to the interval $a < X_i < b$) and setting

$$p(m) = \begin{cases} \left[\Phi\left(\dfrac{b - \overline{X_0}}{\sigma_n}\right) - \Phi\left(\dfrac{a - \overline{X_0}}{\sigma_n}\right)\right]^{-1} \cdot \dfrac{1}{\sqrt{2\pi}\,\sigma_n} \exp\left\{-(m-\overline{X_0})^2/2\sigma_n\right\}, & a < m < b \\ 0, & \text{otherwise} \end{cases} \tag{53}$$

where

$$\overline{X_0} = \frac{1}{t} \sum_{i=-t+1}^{0} X_i \tag{54}$$

$$\sigma_n^2 = \frac{1}{n} \sigma^2 \tag{55}$$

and $\Phi(x)$ is the Gaussian cumulative distribution function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2}\, dx \tag{56}$$

2. Converting Density to Familiar Form

Alternatively, it may be possible to choose $r(\theta)$ in such a way as to convert a probability density into a more familiar form. For example, if the problem consists of learning both the magnitude and the

phase of a complex Gaussian mean (Case 7), the simple reproducing density is listed in Table 2 as:

$$p(a,\phi) = \begin{cases} \dfrac{I_0^{-1}\left[\dfrac{|\bar{X}_n|^2}{4\sigma_n^2}\right]}{2^{1/2}\pi^{3/2}\sigma_n} \exp\left\{-\dfrac{1}{2\sigma_n^2}\left[a^2 - 2a|\bar{X}_n|\cos(\phi-\delta_n) + \frac{1}{2}|\bar{X}_n|^2\right]\right\} & \\ \qquad\qquad a \geq 0,\ -\pi \leq \phi \leq \pi & \\ 0, \qquad\qquad \text{otherwise.} & \end{cases} \tag{57}$$

If $r(a,\phi)$ is taken identically equal to a, then from Eq. (30) (writing the normalizing constant given by the reciprocal of the denominator in Eq. (30) along with the other constant factors involved as a constant K):

$$p(a,\phi) = \begin{cases} Ka \exp\left\{-\dfrac{1}{2\sigma_n^2}\left[a^2 - 2a|\bar{X}_n|\cos(\phi-\delta_n)\right]\right\} & a \geq 0,\ -\pi \leq \phi \leq \pi \\ 0, & \text{otherwise;} \end{cases}$$

$$= \begin{cases} \dfrac{a}{2\pi\sigma_n^2} \exp\left\{-\dfrac{1}{2\sigma_n^2}\left[a^2 - 2a|\bar{X}_n|\cos(\phi-\delta_n) + |\bar{X}_n|^2\right]\right\} & a \geq 0,\ -\pi \leq \phi \leq \pi, \\ 0, & \text{otherwise.} \end{cases} \tag{58}$$

The normalizing constant K was evaluated in the second expression for $p(a,\phi)$ by a procedure suggested in Section A. It was noted that the density depended on its arguments in the same manner as one of the standard densities used in statistical communication theory; in this case the dependence is the same as that of the generalized Rayleigh or Rician density encountered in the study of narrow-band signals in Gaussian noise. Hence, the normalizing constant for the Rician density was used.

3. Other Possibilities

Numerous other reasons for choosing a particular $r(\theta)$ may occur. The form may be determined: by reasoning about physical principles, to agree with experimental results, or in numerous other ways.

4. Computation of Density Needed in Chapter VII

One density of the form in Eq. (52) will be needed in the next chapter. Consider an event E with conditional probability

$$P(E|f) = K_1 \exp\left[\frac{8B^2}{N_o}\left|\int_0^{T_1} x(t)e^{i2\pi ft}dt\right|^2\right] \tag{59}$$

with K_1 a normalizing constant. Assume that f is known to be confined to the interval I for which $f_o \leq f \leq f_1$. To obtain a reproducing density, select a function y(t) which is defined for $-T_o \leq t < 0$ and let

$$p(f) = \begin{cases} K_2 \exp\left[\frac{8B^2}{N_o}\left|\int_{-T_o}^{0} y(t)\, e^{i2\pi ft}dt\right|^2\right], & f_o \leq f \leq f_1 \\ 0, & \text{otherwise.} \end{cases} \tag{60}$$

where K_2 is another normalizing constant. (The normalizing constants are not evaluated in this example since they are complex and are unnecessary for the later analysis.)

The a posteriori density after observing the event E is then

$$p(f|E) = \begin{cases} K_3 \exp\left[\frac{8B^2}{N_o}\left|\int_{-T_o}^{T_1} z(t)e^{i2\pi ft}dt\right|^2\right], & f_o \leq f \leq f_1 \\ 0, & \text{otherwise.} \end{cases} \tag{61}$$

where

$$z(t) = \begin{cases} y(t), & -T_o \leq t < 0 \\ x(t), & 0 \leq t \leq T_1 \end{cases} \tag{62}$$

since x(t) and y(t) are defined on disjoint time intervals.

D. COMPARISON WITH RESULTS OBTAINED BY OTHER INVESTIGATORS

Reproducing-type distributions are used in a number of papers surveyed in the literature. Results obtained in this investigation may be briefly compared with those in a few of the papers in which reproducing-type distributions are used.

1. Abramson, Braverman, Keehn, Bellman, and Mosimann

As already noted, the densities that Abramson, Braverman, Keehn, Bellman and Mosimann [Refs. 7-12] used are the same densities as the simple reproducing-type densities developed in this investigation for the cases considered. The present study has developed methods for generating these densities rather than finding them by an heuristic, or trial-and-error, process.

2. Daly

Daly's problem [Refs. 16 and 17] cannot be solved by the methods developed in the present investigation, since for his densities no sufficient statistics of fixed dimension exist, with the consequence that no reproducing a priori density exists. In fact, the density Eq. (11) that was given in the discussion of a simple case of Daly's problem is a special case of the density

$$p(X|m_1,m_2,\sigma_1^2,\sigma_2^2,P)$$

$$= \frac{P}{\sqrt{2\pi}\,\sigma_1} \exp\left[-\frac{1}{2\sigma_1^2}(X-m_1)^2\right] + \frac{(1-P)}{\sqrt{2\pi}\,\sigma_2} \exp\left[-\frac{1}{2\sigma_2^2}(X-m_2)^2\right] \quad (63)$$

Dynkin [Ref. 19] shows that for the density in Eq. (63) no sufficient statistic of fixed dimension exists if any one of the parameters m_1, m_2, σ_1^2, σ_2^2 or P is unknown.

3. Raiffa and Schlaifer

Raiffa and Schlaifer [Ref. 15] utilize reproducing densities in a large portion of their work on statistical decision theory. Their "natural conjugate" a priori densities are the same form as the simple

reproducing-type densities in the present investigation. Raiffa and Schlaifer do not utilize any specific set of a priori observations to generate the reproducing density, however, merely saying that the density is generated by the kernel of the sufficient statistic for the likelihood [the function $f(t_1, \ldots t_s, \theta)$ in Eq. (27)]. The a priori observations have been utilized in the present work largely as an aid to visualizing the process of generating reproducing-type distributions, and of utilizing the distributions to reflect a priori knowledge.

For small samples at least, a difficulty with the Raiffa-Schlaifer approach lies in ascertaining the number of observations to which the a priori knowledge is equivalent--a problem discussed on pages 62-67 of the work cited [Ref. 15], and also discussed in earlier sections of this report. An example of the difference in methods of interpretation is the case of learning the probability P characterizing a binomial distribution. Raiffa and Schlaifer consider the knowledge reflected in the density Eq. (44) to be equivalent to n_o+2 observations, since Eq. (44) is a valid probability density for $n_o+2 \geq 0$; while in this paper the knowledge is considered to be equivalent to n_o observations. As Raiffa and Schlaifer's equivalent number of observations, n_o+2, approaches zero, the a priori density degenerates into a probability mass function with mass divided between zero and one, a fact that the authors discuss at some length. No matter how many a posteriori observations are then made, the density remains degenerate. In contrast, in the present investigation as the equivalent number of observations n_o approaches zero, Eq. (44) approaches a uniform density (see Table 5)--a much more reasonable result.

Raiffa and Schlaifer also confine their work entirely to simple reproducing densities ("natural conjugate" densities). They make no mention of any other form of densities which may reproduce.

4. Turin

Turin [Ref. 13] utilizes a slight modification of the composite reproducing density Eq. (58) for learning the characteristics of a radio channel. He assumes that a known signal $\mathbf{Y} = (y_1, \ldots y_n)_t$ is transmitted over a channel with amplification a and phase shift ϕ, so

that the received signal is $\mathbf{X} = a\mathbf{Y}e^{j\phi}$. Assuming additive Gaussian noise with mean zero and variance σ^2:

$$p(\mathbf{X}\,|a,\phi,\mathbf{Y}) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum|x_i - a\,y_i\,e^{i\phi}|^2\right] \qquad (64)$$

This equation is the same as the basic equation developed in the present study for the likelihood of a complex Gaussian process with unknown mean (Case 7, Table 2), save for replacing the constant a by the variable ay_i. Following the same procedure used in the present paper in analyzing the complex Gaussian case, and assuming $\mathbf{Y}$ is known, there is obtained for a simple reproducing density on (a,ϕ),

$$p(a,\phi) = \begin{cases} \dfrac{I_0^{-1}\left[\dfrac{R_n^2}{4\sigma_o^2}\right]}{2^{1/2}\pi^{3/2}\sigma_n} \exp\left\{-\dfrac{1}{2\sigma_n^2}\left[a^2 - 2a\,R_n\cos(\phi-\delta_n) + \frac{1}{2}R_n^2\right]\right\} & \\ \qquad\qquad a \geq 0,\ -\pi \leq \phi \leq \pi & \\ 0, \qquad\qquad \text{otherwise.} & \end{cases} \qquad (65)$$

with

$$R_n \triangleq \frac{\left|\sum x_i y_i^*\right|}{\sum |y_i|^2} \qquad (66a)$$

$$\delta_n \triangleq \tan^{-1}\frac{\sum \mathrm{Im}\,(x_i y_i^*)}{\sum \mathrm{Re}\,(x_i y_i^*)} \qquad (66b)$$

$$\sigma_n^2 = \frac{\sigma^2}{\sum |y_i|^2} \qquad (66c)$$

This density reduces to that shown in Table 2 if y_i is taken equal to one for all i.

On the basis of reasoning about the physical process he is considering, Turin picks as a priori density on $(a,\emptyset)$ the Rician density

$$p(a,\emptyset) = \begin{cases} \frac{a}{2\pi\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}[a^2 - 2aR\cos(\emptyset-\delta) + R^2]\right\}, & a \geq 0, \ -\pi \leq \emptyset \leq \pi \\ 0, & \text{otherwise.} \end{cases} \quad (67)$$

which corresponds to Eq. (58) in the same way that Eq. (65) corresponds to the density in Case 7, Table 2. Thus, Turin's density is a composite reproducing density with $r(a,\emptyset)$ equal to a. The analysis developed in the present study shows why Turin's density reproduces itself, and also indicates how alternative reproducing densities which may agree more closely with experiment may be found.

Reproducing distributions are doubtless used elsewhere in the literature. The treatment described in the present paper is more general and thorough than any others that have been found in the literature search, however.

VII. APPLICATIONS

A. PATTERN RECOGNITION, EXPONENTIAL DENSITIES

In the previous work by Abramson, Braverman, and Keehn discussed in Chapter III, reproducing distributions were applied to a pattern-recognition process with learning. Using the methods developed in the present study, it is easily possible to generate reproducing distributions for learning a wide variety of parameters, thus obtaining obvious generalizations of the Abramson, Braverman, and Keehn techniques. One application similar to (but in some respects more complex than) the applications discussed by Abramson, Braverman, and Keehn involves learning the parameters of a non-Gaussian density, and in addition learning the probability of a pattern and using this to adjust a threshold.

Consider a variation of the pattern-recognition problem discussed in Chapter III. It is again desired to find a decision rule minimizing the probability of error in recognition. Equation (8) and the discussion that accompanies it indicate that the optimum decision rule picks the pattern for which $p(X|i)P(i)$ is maximum.

For simplicity assume two possible patterns, designated by the indices 1 and 2. The optimum decision rule is then:

$$d(X) = \begin{cases} 1, & \text{if } \dfrac{p(X|1)}{p(X|2)} \geq \dfrac{P(2)}{P(1)} \\ 2, & \text{otherwise} \end{cases} \tag{68}$$

If it be assumed that $p(X|i)$ is an exponential density with parameter λ_i, Eq. (68) becomes

$$d(X) = \begin{cases} 1, & \text{if } \dfrac{\lambda_1}{\lambda_2} e^{(\lambda_2-\lambda_1)X} \geq \dfrac{P(2)}{P(1)} \\ 0, & \text{otherwise} \end{cases} \tag{69}$$

or

$$d(X) = \begin{cases} 1, & \text{if } (\lambda_2 - \lambda_1)X \geq \ell n \frac{P(2)}{P(1)} + \ell n \frac{\lambda_2}{\lambda_1} \\ 0, & \text{otherwise} \end{cases} \tag{70}$$

When neither the λ_i nor the $P(i)$ is known, the learning procedure developed in this investigation is employed. To learn the λ_i, the simple reproducing density for this case (No. 9 in Table 2) is used. As an a priori density on λ_i the gamma density given by

$$p(\lambda_i) = \frac{C_{oi}}{n_{oi}} (C_{oi}\, \lambda_i)^{n_{oi}}\, e^{-C_{oi}\lambda_i} \tag{71}$$

is assumed. This gives

$$p(X|i) = \int_0^\infty p(X|i,\, \lambda_i)\, p(\lambda_i) d\lambda_i$$

$$= \frac{n_{oi}+1}{C_{oi}} \cdot \frac{1}{(1 + X/C_{oi})^{n_{oi}+2}} \tag{72}$$

It is also desired to learn the probabilities $P(i)$. Letting $P(1)$ equal P and $P(2)$ equal $1-P$, it is seen that P is the parameter characterizing a binomial distribution. Use is again made of a simple reproducing density (in this case No. 1 in Table 2). The number of times each pattern occurs in the "a priori set of observations" is already known; the parameter n_{oi} in Eqs. (71) and (72) corresponds to the number of observations of pattern i. Substituting n_{o1} and n_{o2} for the corresponding parameters r and s in Case 1 of Table 2:

$$p(P) = \frac{\Gamma(n_o+2)}{\Gamma(n_{o1}+1)\Gamma(n_{o2}+1)}\, P^{n_{o1}}\, (1-P)^{n_{o2}} \tag{73}$$

where

$$n_o \triangleq n_{o1} + n_{o2} \tag{74}$$

Then, applying the standard statistical procedure for computing marginal probabilities

$$P(i) = \int_0^1 P(i|P)p(P)\,dP$$

$$= \frac{n_{oi} + 1}{n_o + 2} \tag{75}$$

since $P(1|P) = P$, $P(2|P) = 1-P$. The optimum decision rule then becomes

$$d(X) = \begin{cases} 1, & \text{if } \dfrac{(1 + X/C_{o2})^{n_{o2}+2}}{(1 + X/C_{o1})^{n_{o1}+2}} \geq \dfrac{(n_{o2} + 1)/(n_o + 2)}{(n_{o1} + 1)/(n_o + 2)} \cdot \dfrac{(n_{o2}+1)/C_{o2}}{(n_{o1}+1)/C_{o1}} \\ 2, & \text{otherwise} \end{cases} \tag{76}$$

If n_1 classified learning observations are then taken, with n_{1i} from class i, an "a posteriori decision rule" of identical form results except for replacing n_{oi} by $n_{oi}+n_{1i} \triangleq n_{ti}$, n_o by $n_o+n_1 \triangleq n_t$, and C_{oi} by $C_{oi}+C_{1i} \triangleq C_{ti}$ (with C_{1i} the sum of the X_j that correspond to the i^{th} pattern). The optimum decision rule after n_1 observations is:

$$d_{n_1}(X) = \begin{cases} 1, & \text{if } \dfrac{\left(1 + \dfrac{X}{C_{t2}}\right)^{n_{o2}+2}}{\left(1 + \dfrac{X}{C_{t1}}\right)^{n_{o1}+2}} \geq \dfrac{(n_{o2}+1)/(n_o+2)}{(n_{t1}+1)/(n_t+2)} \cdot \dfrac{(n_{t2}+1)/C_{t2}}{(n_{t1}+1)/C_{t1}} \\ 2, & \text{otherwise} \end{cases} \tag{77}$$

Since $(n_{ti}+1)/(n_t+2)$ is an estimate of $P(i)$, it is designated by $\hat{P}(i)$. Similarly, $(n_{ti}+1)/C_{ti}$ is designated by $\hat{\lambda}_i$ since it is an estimate of the parameter λ_i. Taking logarithms in Eq. (77):

$$d_{n_1}(X) = \begin{cases} 1, \quad \text{if} \quad (n_{t2}+2)\left[\frac{X}{C_{t2}} - \frac{1}{2}\left(\frac{X}{C_{t2}}\right)^2 + \ldots\right] \\ \qquad -(n_{t1}+2)\left[\frac{X}{C_{t1}} - \frac{1}{2}\left(\frac{X}{C_{t1}}\right)^2 \ldots\right] \geq \ell n \frac{\hat{P}(2)}{\hat{P}(1)} + \ell n \frac{\hat{\lambda}_2}{\hat{\lambda}_1} \\ 2, \quad \text{otherwise.} \end{cases} \tag{78}$$

The quantity X/C_{ti} can normally be expected to be of the order $1/n_{ti}$. Hence, after a few observations, the first term in the expansion of the logarithm becomes predominant and higher-order terms can be neglected. After a few observations it is also possible to neglect the difference between $n_{ti}+2$ and $n_{ti}+1$. After a few observations, then, the optimum decision rule given by Eq. (77) is closely approximated by the decision rule

$$d'_{n_1}(X) = \begin{cases} 1, \quad \text{if} \quad (\hat{\lambda}_2 - \hat{\lambda}_1)X \geq \ell n \frac{\hat{P}(2)}{\hat{P}(1)} + \ell n \frac{\hat{\lambda}_2}{\hat{\lambda}_1} \\ 0, \quad \text{otherwise.} \end{cases} \tag{79}$$

This is of the same form as Eq. (70). Hence, it may be concluded that after a few observations are taken, the optimum decision rule is closely approximated by a rule that is of the established form for known statistics, but which utilizes estimates of the parameters in place of the parameters themselves.

The approximate decision rule derived in Eq. (79) can be implemented as shown in Fig. 6 by a device of the form which would be applicable with known parameters, but with variable components.

Since the $\hat{\lambda}_i$ may take on any positive values and the $\hat{P}(i)$ any values between zero and one, the Bayes' decision rules computed from

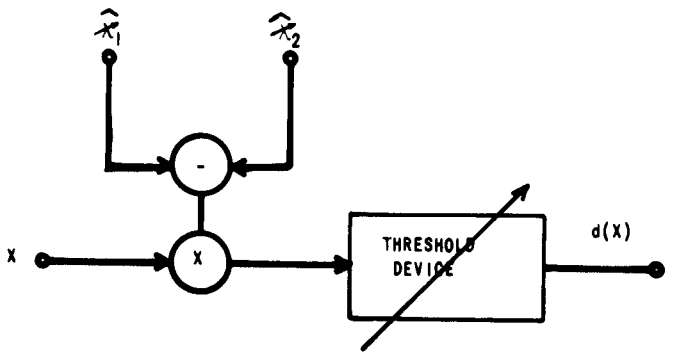

FIG. 6. PATTERN CLASSIFIER FOR EXPONENTIAL DENSITIES.

Eq. (79) can assign all X's below any real-number threshold to class 1 and those above the threshold to class 2; or vice versa. In other words, any non-randomized decision rule based on a single threshold is a possible Bayes' rule.

The estimate of each of the parameters used in Eq. (79) converges with probability one to the true value of the parameter. Hence, the limiting form of the decision rule given in Eq. (79) is identical to the rule that would be used if all the parameters were known. This again could be any non-randomized decision rule based on a single threshold.

B. FINDING EXPECTATION OF A RANDOM VARIABLE

Another class of problems for which reproducing densities are applicable is that of finding the expectation of a random variable. More precisely, reproducing densities are useful in cases where a probability density is required that will adequately represent a priori information and at the same time allow the expected value of a non-negative random variable to be expressed in a simple form. This type of problem may be illustrated by considering the problem of detecting a cosine of unknown frequency.* Two possible hypotheses are assumed:

*This example was suggested and first worked out by Professor Norman Abramson, Stanford University.

$$\left.\begin{aligned} H_1: &\quad X(t) = S(t) + N(t) \\ H_2: &\quad X(t) = N(t) \end{aligned}\right\} \tag{80}$$

where

$$S(t) = a \cos(\omega t + \phi), \qquad \omega = 2\pi f \tag{81}$$

and $N(t)$ is white noise, or noise with a flat spectrum $S_n(f) \equiv N_o/2$ (at least over the frequency range $f_o \leq f \leq f_1$).

It is assumed that the parameters a, ϕ, and f (or ω) are all unknown, although the following are known: (1) that ϕ is uniformly distributed over the range $0 \leq \phi \leq 2\pi$; (2) that a is Rayleigh-distributed with parameter A^2; and (3) that f is restricted to the frequency range $f_o \leq f \leq f_1$. It is desired to use a likelihood ratio test, comparing

$$\ell(X) = \frac{p(X|H_1)}{p(X|H_2)} \tag{82}$$

with some threshold.

If a, ϕ, and f were known, the likelihood of a sample $X(t)$, $0 \leq t \leq T_1$, would be

$$\ell(X|a,\phi,f) = \exp\left\{-\frac{2a^2}{N_o}\int_0^{T_1} \cos^2(\omega t+\phi)dt\right\} \exp\left\{\frac{4a}{N_o}\int_0^{T_1} X(t)\cos(\omega t+\phi)\,dt\right\}$$

$$\approx \exp\left\{-\frac{a^2 T_1}{N_o}\right\} \exp\left\{\frac{4a}{N_o}\int_0^{T_1} X(t)\cos(\omega t+\phi)\,dt\right\} \tag{83}$$

In writing the last form of the equation it has been assumed that T_1 is large in comparison with $1/f_o$, so that the integral of the cosine-squared term is approximately $\frac{1}{2}$ regardless of ω or ϕ.

It is shown in the Appendix that, with the likelihood given in Eq. (83) and with the probability densities assumed for ϕ and a,

$$\ell(X|f) = K_1 \exp\left\{\frac{8B^2}{N_o}\left|\int_0^{T_1} X(t)\, e^{i\omega t}\right|^2\right\} \tag{84}$$

with

$$B^2 = \left[\frac{A^2}{2A^2T_1+N_o}\right] \tag{85}$$

It is desired to find a probability density $p(f)$ which will give a reasonably simple form for $\ell(X)$ and at the same time accurately reflect any information that is available about f. Such a density is obtained by following the same process that was used in finding reproducing densities. Although $\ell(X|f)$ is not a probability density, it is non-negative. If $\ell(X|f)$ were normalized to integrate to one, it would satisfy the formal requirements for a probability density. This is the same procedure used to derive reproducing-type densities from likelihood functions; this suggests deriving a density for f in the same manner. Such a density was derived in Chapter VI, Section C, and is given by Eq. (60). Utilizing the density in Eq. (60) for f gives

$$\ell(X) = K_4 \int_{f_o}^{f_1} \exp\left\{\frac{8B^2}{N_o}\left|\int_{-T_o}^{T_1} Z(t)\, e^{i\omega t}\, dt\right|^2 df\right\} \tag{86}$$

with

$$Z(t) = \begin{cases} Y(t), & -T_o \le t < 0 \\ X(t), & 0 \le t \le t_1 \end{cases} \tag{87}$$

and K_4 a new constant that may be absorbed into the threshold for the likelihood-ratio test.

Without specifying $X(t)$ and $Y(t)$ more definitely, the integrals in Eq. (86) cannot be evaluated. However, the following points may be noted. If T_o is small, the frequency information in Eq. (86) is primarily determined by $X(t)$; if T_o is large, the information

is primarily determined by $Y(t)$. Hence, T_o is a measure of confidence in the a priori information. Also, by proper choice of $Y(t)$, $p(f)$ can be caused to peak around any desired frequency band. The density given by Eq. (60) appears to be the only one yet found with these properties, which are important for this application.

C. ESTIMATING A PARAMETER WITH NO A PRIORI INFORMATION

1. Bayes Estimates

In order to compute the Bayes estimate of a parameter it is necessary to specify an a priori probability distribution for the parameter. If no information about this distribution is available, and if no reason is known for favoring some values of the parameter, a uniform a priori probability distribution is the logical assumption. It is only possible to assume a uniform distribution if the range of the parameter is of finite Lebesgue measure, however.*

The techniques developed in this investigation can be used to eliminate this difficulty. To illustrate the procedure, assume that it is desired to estimate a parameter ω, and that a squared-error loss function is involved:

$$L(\omega, \hat{\omega}) = (\omega - \hat{\omega})^2 \tag{88}$$

where $\hat{\omega}$ is the available estimate of ω. It is well known [Ref. 20] that the Bayes estimate for this case is the a posteriori expected value of ω, or

$$\hat{\omega}(X) = \int \omega p(\omega|X)\, d\omega \tag{89}$$

with X the observation that is being utilized to estimate ω.

The function $p(\omega|X)$ is an a posteriori density function, of the form that has been studied in this investigation. If it is desired to

*As mentioned earlier, uniform densities over ranges of infinite measure are allowed in the theory developed by Rényi [Ref. 21].

approximate the form that the Bayes estimate would take with a uniform a priori density over ω, a reproducing-type a priori density on ω can be assumed, then the size of the set of a priori observations can be allowed to approach zero. It has been shown that the reproducing density then approaches a uniform density. At the same time, however, the a posteriori density $p(\omega|X)$ approaches $\hat{p}(\omega|X)$, if this latter density is defined. (This may be seen by examining the form of Eq. (31) as the size of the set of a priori observations approaches zero, with $r(\theta)$ set equal to a constant.)

The following result is thus obtained: The limiting form of the Bayes estimate of ω as the a priori density on ω approaches uniformity is given by

$$\hat{\omega}(X) = \int \omega\, \hat{p}(\omega|X)\, d\omega \qquad (90)$$

where $\hat{p}(\omega|X)$ is an "experimental" probability density of the form defined in Eq. (20).

If the estimate is based on a sequence of measurements $\{X_1, \ldots X_n\}$, the same result is obtained, but with $\hat{p}(\omega|X)$ replaced by $\hat{p}(\omega|X_1, \ldots X_n)$. The Bayes estimates are given by the mean values listed in Table 3 for the cases studied in this investigation; no distinction was made between a priori and a posteriori observations in making up this table.

The derivation given above is based on the assumption of a squared-error loss function. Bayes estimates with other loss functions, if they can be evaluated, are also given in terms of a posteriori densities. Estimates with no a priori knowledge would be obtained in a manner analogous to that just described.

2. Maximum-Likelihood Estimates

Maximum-likelihood estimates are often used instead of Bayes estimates if no a priori information is available. The techniques discussed in this report can also be used to simplify the procedure for obtaining maximum-likelihood estimates. These estimates correspond to the mode of the likelihood function, or the value of ω for which the

likelihood function is maximum. This mode is also the mode of the "experimental portion" of the a posteriori density, since this portion is simply a normalized version of the likelihood function. If the "experimental portion" of the density is of fixed form, the mode can normally be expressed as a fixed function of the parameters characterizing the density. Expressing the parameters characterizing the density in terms of the sufficient statistics for the observations, and the mode in terms of these parameters, a recursive method for computing the maximum likelihood estimates is obtained. The maximum-likelihood estimates may in this manner be expressed as explicit functions of the observations.

The two methods discussed above for estimating parameters when no a priori information is available are not equivalent, although the difference is negligible for large numbers of learning observations. For example, in estimating the parameter P of a binomial distribution, the maximum likelihood and Bayes estimates are r/n and $(r+1)/(n+2)$ respectively, while in estimating the covariance matrix of a Gaussian density, the corresponding estimates are V_n/n and $V_n/(n+d+1)$.

VIII. SUMMARY AND CONCLUSIONS

A model has been developed for a learning technique capable of utilizing and evaluating statistical information relating to a physical system or process. Characteristics of the technique are as follows:

A. BASIC ASSUMPTIONS

1. A body of statistics is available, or can be obtained, about the system or process under study.
2. In these statistics there are one or more parameters, denoted by θ, whose values are unknown.
3. Each unknown parameter θ can be treated as a random variable having a probability density $p(\theta)$ over the range of its possible values. (The expedient of treating θ in this manner is typical of the "Bayesian" approach to probability theory.)
4. A priori information is available to aid in choosing the probability density $p(\theta)$. This a priori information can involve information gained from a knowledge of the physical principles involved in the process, information gained from experience, or information gained in other ways.
5. It is possible to perform experiments on the system, yielding sets of learning observations $\Lambda_1, \ldots \Lambda_n$.
6. The likelihood of each set of learning observations Λ_i is known as a function of θ, and is designated as $p(\Lambda_i|\theta)$. (When viewed as a function of θ for fixed Λ_i, $p(\Lambda_i|\theta)$ is called a likelihood function; when viewed as a function of Λ_i for fixed θ, $p(\Lambda_i|\theta)$ is called a conditional-probability-density function.)
7. The learning observations $\Lambda_1, \ldots \Lambda_n$ are used only to gain knowledge about θ, and do not influence the values of θ.
8. A random variable Z may be selected to represent some desired criterion of system performance, such as the fraction of the time the system makes an error, or some other error function.

9. The excellence of system performance may be judged by the statistical expectation of Z, E[Z], where

$$E[Z] = \int E[Z|\theta]\, p(\theta)\, d\theta \qquad (1)$$

10. In the above equation, $E[Z|\theta]$ is the conditional expectation of Z given θ, expressed as a function of θ, and is independent of $\Lambda_1, \ldots \Lambda_n$. $E[Z|\theta]$ is assumed to be known a priori.

B. DEVELOPMENT OF BASIC LEARNING MODEL

1. Apply "Bayes' rule" to obtain

$$p(\theta|\Lambda_1) = \frac{p(\Lambda_1|\theta)\, p(\theta)}{\int p(\Lambda_1|\theta)\, p(\theta)\, d\theta} \qquad (2)$$

where

$p(\theta|\Lambda_1)$ = a posteriori probability density of θ
= probability density of θ evaluated in the the light of the set of learning observations Λ_1,

and also

$p(\theta)$ = a priori probability density of θ,
$p(\Lambda_1|\theta)$ = likelihood of the learning observations Λ_1.

2. Then Eq. (1) becomes

$$E[Z|\Lambda_1] = \int E[Z|\theta]\, p(\theta|\Lambda_1)\, d\theta \qquad (3)$$

where

$E[Z|\Lambda_1]$ = statistical expection of Z, in the light of the learning observations Λ_1,

and

$E[Z|\theta]$ = conditional expectation of Z given θ, expressed as a function of θ.

3. An additional set of learning observations Λ_2 is obtained and Bayes' rule, Eq. (2) is again applied to obtain:

$$p(\theta|\Lambda_1,\Lambda_2) = \frac{p(\Lambda_2|\theta,\Lambda_1)p(\theta|\Lambda_1)}{\int p(\Lambda_2|\theta,\Lambda_1)p(\theta|\Lambda_1)\,d\theta} \tag{4}$$

4. The process is repeated to yield, eventually,

$$p(\theta|\Lambda_1, \ldots \Lambda_n) = \frac{p(\Lambda_n|\theta,\Lambda_1, \ldots \Lambda_{n-1})p(\theta|\Lambda_1,\ldots\Lambda_{n-1})}{\int p(\Lambda_n|\theta,\Lambda_1, \ldots \Lambda_{n-1})p(\theta|\Lambda_1, \ldots \Lambda_{n-1})\,d\theta} \tag{5}$$

where

$p(\theta|\Lambda_1, \ldots \Lambda_n)$ = the a posteriori probability density of θ in the light of the first n sets of learning observations;

$p(\Lambda_n|\theta,\Lambda_1, \ldots \Lambda_{n-1})$ = the likelihood of the n^{th} set of observations given the first n-1 sets of observations.

5. If it be assumed that the sets of learning observations are conditionally independent given θ, Eq. (5) may be simplified to:

$$p(\theta|\Lambda_1, \ldots \Lambda_n) = \frac{p(\Lambda_n|\theta)p(\theta|\Lambda_1, \ldots \Lambda_{n-1})}{\int p(\Lambda_n|\theta)p(\theta|\Lambda_1, \ldots \Lambda_{n-1})\,d\theta} \tag{6}$$

6. Using Eq. (6) above (or Eq. (5)), Eq. (3) is expanded to give:

$$E[Z|\Lambda_1, \ldots \Lambda_n] = \int E[Z|\theta]p(\theta|\Lambda_1, \ldots \Lambda_n)\,d\theta \tag{7}$$

7. Equations (6) and (7) above form the basis for the learning model illustrated in Fig. 1 of the report.

C. CONDITIONS FOR FEASIBILITY OF THE LEARNING PROCESS

1. The learning technique described above may be considered to be a practical learning process if:
   a. The true values of the unknown parameters are eventually learned, at least in the limit as the number of learning observations approaches infinity. This condition may be considered to be met if, as the number of learning observations approaches infinity, the a posteriori density $p(\theta|\Lambda_1, \ldots \Lambda_n)$ approaches a Dirac delta function at the true values of the unknown parameters.
   b. The form of the learning process does not change as additional observations are taken. This condition may be considered to be met if the probability distributions on θ are <u>reproducing</u> in nature--i.e., if the a posteriori and a priori distributions are of the same form under Bayes' rule. If the distributions are reproducing, the learning process simply involves computation of new parameters for the densities at each stage of the process, neither the number nor the type of computations changing.
2. Condition (a) is fulfilled if it is possible to compute the true value of θ from an infinite sequence of learning observations; and this true value is not ruled out by $p(\theta)$, the a priori probability distribution assumed for θ. It is shown in the report that these conditions are met by most probability distributions of practical significance, even by some distributions of such form that condition (b) cannot be met. Thus, the learning process developed in this report should be valid for most practical cases, provided condition (b) is also fulfilled.
3. In order to determine whether the a priori $p(\theta)$ assumed is reproducing or not [condition (b)] a technique has been developed whereby the expression for the a posteriori density is factorized as follows:

$$p(\theta|\Lambda_1, \ldots \Lambda_n) = \hat{p}(\theta|\Lambda_1, \ldots \Lambda_n) \cdot \frac{p(\theta)}{\hat{E}[p(\theta)|\Lambda_1, \ldots \Lambda_n]} \quad (19)$$

wherein

$$\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n) = \frac{p(\Lambda_1, \ldots \Lambda_n|\theta)}{\int p(\Lambda_1, \ldots \Lambda_n|\theta)\, d\theta} \qquad (20)$$

= "experimental portion" of a posteriori density (depends only on the observations),

$\hat{E}[p(\theta)|\Lambda_1, \ldots \Lambda_n]$ = statistical expectation of $p(\theta)$ taken with respect to $\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n)$.

The likelihood function $p(\Lambda_1, \ldots \Lambda_n|\theta)$ used to generate $\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n)$ is assumed to be an integrable, non-negative function of θ; the "experimental portion" of the a posteriori density is a normalized version of the likelihood.

4. It is shown in the report that (at least after a large number of learning observations) the behavior of the a posteriori density $p(\theta|\Lambda_1, \ldots \Lambda_n)$ is primarily determined by the "experimental portion" $\hat{p}(\theta|\Lambda_1, \ldots \Lambda_n)$, see Eq. (19) above. Conditions for the "experimental portion" to be reproducing are analyzed in the report. It is shown that the "experimental portion" of the a posteriori density is reproducing if and only if the learning observations are such that a sufficient statistic for θ of fixed dimension exists.

5. It is possible to find an a priori $p(\theta)$ that is reproducing if and only if the "experimental portion" of the a posteriori density is reproducing, i.e., if and only if a sufficient statistic for θ of fixed dimension exists. Any reproducing $p(\theta)$ that exists may be generated by multiplying a function of the form of the likelihood $p(\Lambda_1, \ldots \Lambda_n|\theta)$ by an arbitrary non-negative function of θ and then normalizing.

6. If a sufficient statistic for θ of fixed dimension exists, the a posteriori densities $p(\theta|\Lambda_1)$, $p(\theta|\Lambda_1,\Lambda_2)$, ... become reproducing <u>after</u> the first observation has been utilized (occasionally after the first few observations have been utilized). Hence, if there is no objection to one reprogramming of the learning system

after the first set of learning observations, the learning techniques described herein can be applied regardless of what a priori density $p(\theta)$ is used, provided a sufficient statistic of fixed dimension exists.

7. Since this is the case, the use of reproducing-type a priori densities may in many cases afford little if any simplification in the computations involved. Non-reproducing densities might be preferred if they resulted in a faster rate of convergence to a delta function of the a posteriori probability densities. It is shown, however, that little if any increase in rate of convergence can be obtained by using non-reproducing densities, if the a priori densities are approximately the same width.
8. The results can be generalized to apply to the case where the learning observations $\Lambda_1, \ldots \Lambda_n$ are not conditionally independent given θ; however, in this case the form of the learning system may depend on the state of the system derived from the previous observations.

D. EXAMPLES OF REPRODUCING-TYPE DENSITIES

1. Two classes of reproducing-type densities are considered:
   a. Simple reproducing-type densities are densities identical in form with the "experimental portion" of the a posteriori density. Such densities may be generated by picking the "a priori observations" $\{\Lambda_{-m}, \ldots \Lambda_0\}$, then normalizing the likelihood for these observations as in Eq. (20) above.
   b. Composite reproducing-type densities are simple reproducing-type densities multiplied by another function of θ and then normalized; i.e., composite reproducing-type densities are of the form

$$p(\theta) = \frac{\hat{p}(\theta|\Lambda_{-m}, \ldots \Lambda_0)\, r(\theta)}{\int \hat{p}(\theta|\Lambda_{-m}, \ldots \Lambda_0)\, r(\theta)\, d\theta} \tag{30}$$

   where $r(\theta)$ is a non-negative, integrable function of θ.

2. Tables 1 through 5 list a number of simple reproducing type probability densities, with many of their parameters and properties. Methods of utilizing the densities to represent a priori knowledge are discussed; the limiting forms of the densities as the number of observations becomes very small or very large are also given.
3. Two important classes of composite reproducing-type densities are discussed. The first class is applicable when the parameter θ is known to lie within a certain range, but no parts of this range are to be preferred over others. The second class arises from the possibility of choosing $r(\theta)$ to convert an unfamiliar probability density into a more familiar form. Numerous other types of composite reproducing-type densities are possible.

E. APPLICATIONS

1. As long as a sufficient statistic of fixed dimension exists, the techniques herein developed are applicable to a wide variety of problems such as pattern recognition with incomplete knowledge of the statistics involved, finding a probability density that simplifies taking the expectation of a non-negative random variable, or estimating a parameter when no a priori information is available. The problems include some for which the learning model developed in this paper is not applicable.
2. The chief requirement for application of the technique is the existence of a sufficient statistic of fixed dimension. Dynkin [Ref. 19] has made a general study of the conditions under which sufficient statistics of fixed dimension exist, and of methods for finding them. Sufficient statistics of fixed dimension appear to exist for most of the simpler probability laws normally encountered, and for some of the more complex ones.

IX. RECOMMENDATIONS FOR FURTHER WORK

Although the results of this investigation give solutions to a number of problems in the field of machine learning, they open up a number of new problems. These problems include finding methods for extending the present theory and finding methods for tying the present theory in with other results in the machine-learning area. Some of these problems are indicated below.

A. PROBLEMS SUGGESTED

1. Procedure When Sufficient Statistics do not Exist

Much of the work on the theory of communication systems involves analyzing complex systems. The probability laws encountered in studying the more complex systems (and some of the simpler ones) are often of forms for which no simple, sufficient statistics exist. In these cases the theory developed in this paper is not directly applicable.

One of the chief problems to be investigated is finding how to proceed when no simple, sufficient statistic exists. A possible approach would be to use a statistic that is not sufficient, but that is of fixed dimension and in some sense "efficient." If this approach is to be used, some method of comparing possible statistics is needed. A criterion might be based on Kullback's information integral or divergence [Refs. 22, 23], which are maximum if and only if based on a sufficient statistic.

2. Effect of Taking Expectation of Performance Criterion

The analysis herein has been confined almost exclusively to the computation of the probability densities $p(\theta|\Lambda_1, \ldots \Lambda_n)$. In actual applications, these probability densities would normally be used to take the expectation of some random variable (see the final stage in Fig. 1). The forms that this final stage of the computation might take and the effects of these forms on the learning process should be investigated. The chief result along these lines in this investigation is the proof that the limiting form for the total computation is the

form that would have been obtained if the unknown parameters had been known (Corollary, Theorem I, Chapter IV).

3. Rate of Convergence

Little work has been done on investigating the rate at which the probability densities converge to their limiting (delta function) form. Since it has been shown that the convergence properties are determined largely by the "experimental portion" of the a posteriori density, and since this portion of the density is a normalized likelihood, some of the techniques employed in the study of maximum likelihood estimates may be useful here.

4. Applications

The material presented in this paper has only begun to scratch the surface of the possible applications of the techniques that have been developed. The problem has been formulated in a general enough manner to indicate that there is a wide variety of possible applications; however, a great deal of work on specific applications remains to be done.

5. Information-Theory Properties

The probability densities examined in this paper appear to have some interesting information-theory properties. These aspects have not been investigated as yet. It may be possible to tie the theory developed in this paper in with some models for learning processes that are based on such information-theory concepts as entropy [Refs. 24, 25].

6. Effects of Errors

If an error is made in the type of likelihood function, $p(\Lambda_1|\theta)$ assumed, the results are unpredictable. (This does not contradict the proof that the limiting form of the a posteriori density is independent of the a priori density, as in this case $p(\Lambda_1|\theta)$ was not in error.) The form that the a posteriori density will take in the limit can be predicted in any particular case. For example, if it were assumed that the observations were generated by a one-dimensional Gaussian process with the density having known variance and unknown

mean, whereas the input observations were actually generated by an exponential process, the sample average would be used as an estimate of the mean while this sample average was actually converging to $1/\lambda_o$ (see Tables 2 and 4). How accurately the resulting probability distribution would fit the data is not clear. This question would be worth investigating, as would a more general analysis of the effects of errors.

7. Several Possible Likelihood Functions

In certain cases it might be known that the likelihood function took one of several possible forms, such as Gaussian, Rayleigh, or exponential, but the precise one of these forms applicable might not be known. In such cases an approach assuming a number of possible forms for the likelihood function is possible, weighting each of these hypotheses by a factor similar to Watanabe's credibility measure [Ref. 26], and adjusting the weights as observations are taken may be feasible. A similar problem has been investigated by Magill [Ref. 27] in developing techniques to predict which of a known set of possible Gaussian signals is being observed, and at the same time predict the value of the signal.

B. SUMMARY

In summary, a fairly general theory has been developed, which appears to have wide applicability; however, much additional work on extending the theory, tying it in with other theories, and applying it to specific cases remains to be done.

APPENDIX

DETAILED COMPUTING PROCEDURES

This appendix describes the detailed procedures used in computing the densities, limits, and so on in Tables 1 through 5: it includes also a special computation for the expectation of a cosine of unknown frequency for Chapter VII.

A. COMPUTATION OF REPRODUCING DENSITIES

It is desired to compute the forms of the simple reproducing densities listed in Table 2, plus the simple reproducing density for the Gaussian case with both $\mathbf{M}$ and $\mathbf{K}$ unknown.

The first density, the beta density for learning P for a binomial distribution, was computed in the main text. The computation simply involves normalizing the likelihood function in the first column of Table 2. This can be done either by integration or by comparing with standard densities as discussed in the text. A similar procedure is followed in all the cases in Table 2.

The second and third densities in Table 2 are generalizations of the first and need no discussion. The derivation of the fourth, a gamma density for learning the parameter α for a Poisson distribution, is also straightforward. It is simplified slightly if the likelihood is rewritten as

$$p(n,\tau|\alpha) = K(n,\tau)\ \alpha^{n}\ e^{-\alpha\tau} \qquad (A.1)$$

and only the part depending on α is considered in normalizing.

The fifth density, Gaussian for learning a Gaussian mean, is derived in a similar manner. The computation is simplified by completing the square in the exponent of the likelihood, using

$$\sum (\mathbf{x}_i - \mathbf{M})_t\, \mathbf{K}^{-1} (\mathbf{x}_i - \mathbf{M}) = n(\bar{\mathbf{x}}_n - \mathbf{M})_t\, \mathbf{K}^{-1} (\bar{\mathbf{x}}_n - \mathbf{M}) + \sum (\mathbf{x}_i - \bar{\mathbf{x}}_n)_t\, \mathbf{K}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_n). \qquad (A.2)$$

The likelihood is then rewritten as

$$p(\mathbf{X}_1, \ldots \mathbf{X}_n | \mathbf{M}) = K(\mathbf{X}_1, \ldots \mathbf{X}_n) \exp\left[-\tfrac{1}{2}(\bar{\mathbf{X}}_n - \mathbf{M})_t \mathbf{K}_n^{-1}(\bar{\mathbf{X}}_n - \mathbf{M})\right] \tag{A.3}$$

proceeding thereafter as in the Poisson case.

The Wishart density for learning an unknown covariance matrix (Case 6) is derived in a similar manner, utilizing the identity

$$\operatorname{tr} \mathbf{V}_n \mathbf{K}^{-1} = \sum_{i=1}^{n} (\mathbf{X}_i - \mathbf{M})_t \mathbf{K}^{-1} (\mathbf{X}_i - \mathbf{M}) \tag{A.4}$$

to show that the two forms of the likelihood in the fifth and sixth cases of Table 2 are equivalent. In this case, comparing the manner in which the likelihood depends on $\mathbf{K}^{-1}$ with the manner in which the Wishart density depends on $\mathbf{V}_n$ is much simpler than integration as a method of obtaining the normalizing constant. See Chapter VI, Section A for a discussion of this procedure.

If both $\mathbf{M}$ and $\mathbf{K}^{-1}$ are unknown, $p(\mathbf{X}_1, \ldots \mathbf{X}_n | \mathbf{M}, \mathbf{K}^{-1})$ is rewritten as

$$p(\mathbf{X}_1, \ldots \mathbf{X}_n | \mathbf{M}, \mathbf{K}^{-1}) = [(2\pi)^d |\mathbf{K}|]^{-(n-1)/2} \exp\left[-\tfrac{1}{2} \operatorname{tr} \mathbf{V}_n^* \mathbf{K}^{-1}\right]$$
$$\cdot [(2\pi)^d |\mathbf{K}|]^{-1/2} \exp\left[-\tfrac{1}{2}(\bar{\mathbf{X}}_n - \mathbf{M})_t \mathbf{K}_n^{-1}(\bar{\mathbf{X}}_n - \mathbf{M})\right] \tag{A.5}$$

with

$$\mathbf{V}_n^* \triangleq \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}}_n)(\mathbf{X}_i - \bar{\mathbf{X}}_n)_t, \tag{A.6}$$

and the other terms defined as before.

The second factor in Eq. (A.5) depends on its parameter in the manner in which a Gaussian density depends on its argument, while the

first factor depends on its parameter in the manner of a Wishart density. This suggests as a normalized density*

$$p(\mathbf{M}, \mathbf{K}^{-1} | \mathbf{X}_1, \ldots \mathbf{X}_n) = \frac{|\mathbf{V}_n^*|^{-(n+d)/2} \, |\mathbf{K}|^{-(n-1)/2} \exp\left[-\frac{1}{2} \operatorname{tr} \mathbf{V}_n^* \mathbf{K}^{-1}\right]}{2^{\frac{1}{2}d(n+d)} \, \pi^{d(d-1)/4} \prod_{\alpha=0}^{d-1} \Gamma\left(\frac{n+d-\alpha}{2}\right)}$$

$$\cdot \left\{(2\pi)^d |\mathbf{K}_n|\right\}^{-1/2} \exp\left\{-\frac{1}{2} (\mathbf{M} - \bar{\mathbf{X}}_n)_t \mathbf{K}_n^{-1} (\mathbf{M} - \bar{\mathbf{X}}_1)\right\} \quad (A.7)$$

The normalization in Eq. (A.7) can be checked by integrating first over $\mathbf{M}$, then over $\mathbf{K}^{-1}$. The first integration gives a Wishart density as a marginal density; the integral of this Wishart density is then unity as it should be.

$\mathbf{V}_n^*$ is the only parameter that has been encountered in a simple reproducing-type density for which a recurrence relation for computing the new value of the parameter from its old value and the learning observations is not obvious. A simple recurrence relation exists, however, as follows:

$$\mathbf{V}_n^* = \mathbf{V}_{n-1}^* + \frac{n-1}{n} [(\mathbf{X}_n - \bar{\mathbf{X}}_{n-1})(\mathbf{X}_n - \bar{\mathbf{X}}_{n-1})_t] \quad (A.8)$$

To derive the density for learning the magnitude and phase of a complex Gaussian mean (Case 7), the portion of the exponent in the likelihood depending on a and ϕ is first rewritten as follows:

$$-2a \sum |X_i| \cos(\phi + \alpha_i) + \sum a^2 = -2na|\bar{X}_n| \cos(\phi + \delta_n) + na^2 \quad (A.9)$$

with $|\bar{X}_n|$ and δ_n defined in Table 2. The normalization is then accomplished by computing

*The density in Eq. (A.7) is not included in Table 2. It is the simple reproducing-type density for learning both $\mathbf{M}$ and $\mathbf{K}^{-1}$ and is the density utilized by Keehn for this purpose [Ref. 10].

$$\int_0^{\infty}\int_{-\pi}^{\pi} \exp\left\{-\frac{1}{2\sigma_n^2}\left[a^2 - 2a|\bar{X}_n| \cos(\phi + \delta_n)\right]\right\} \qquad (A.10)$$

utilizing some of the properties of Bessel functions [Ref. 28].

The final three cases in Table 2 are straightforward. The densities for the exponential and Rayleigh cases may be normalized by comparing with the gamma density; but the density for the rectangular distribution must be normalized by integration.

B. COMPUTATION OF MOMENTS

The moments given in Table 3 were arrived at as follows: beta, Dirichlet, gamma, Gaussian, and Wishart densities are standard forms with moments already tabulated [Refs. 29 - 31]. Hence, in this appendix it is merely necessary to compute the means and the variances for the two cases (Cases 7 and 10) where the simple reproducing-type densities are not standard forms.

The expectation of a, the magnitude of the complex mean of a Gaussian density (Case 7), is given by

$$E[a] = \frac{I_0^{-1}\left[\frac{|\bar{X}_n|^2}{4\sigma_n^2}\right]}{2^{1/2}\pi^{3/2}\sigma_n} \exp\left[-|\bar{X}_n|^2/4\sigma_n^2\right]$$

$$\cdot \int_{-\pi}^{\pi}\int_0^{\infty} a \exp\left\{-\frac{1}{2\sigma_n^2}\left[a^2 - 2a|\bar{X}_n| \cos(\phi - \delta_n)\right]\right\} d\phi\, da \qquad (A.11)$$

It is known that the integral of the Rician density is unity, or

$$\frac{\exp\left[-|\bar{X}_n|^2/2\sigma_n^2\right]}{2\pi\sigma_n^2}\int_{-\pi}^{\pi}\int_0^{\infty} a \exp\left\{-\frac{1}{2\sigma_n^2}\left[a^2 - 2a|\bar{X}_n| \cos(\phi - \delta_n)\right]\right\} d\phi\, da = 1 \qquad (A.12)$$

Comparing Eqs. (A.12) and (A.11) it is found that:

$$E[a] = \left(\frac{2}{\pi}\right)^{1/2} \sigma_n \exp\left[|\bar{X}_n|^2/4\sigma_n^2\right] I_0^{-1}\left[|\bar{X}_n|^2/4\sigma_n^2\right] \qquad (A.13)$$

To obtain the variance, the same procedure is followed, using the fact [Ref. 32] that the first moment of the Rician density is given by

$$\int_{-\pi}^{\pi}\int_{0}^{\infty} a^2 \exp\left\{-\frac{1}{2\sigma_n^2}\left[a^2 - 2a|\bar{X}_n|\cos(\phi - \delta_n) + |\bar{X}_n|^2\right]\right\} d\phi\, da$$

$$= \left(\frac{\pi}{2}\right)^{1/2} \sigma_n \exp\left[-|\bar{X}_n|^2/4\sigma_n^2\right]\left[\left(1 + \frac{|\bar{X}_n|^2}{2\sigma_n^2}\right) I_0\left(\frac{|\bar{X}_n|^2}{4\sigma_n^2}\right) + \frac{|X_n|^2}{2\sigma_n^2} I_1\left(\frac{|\bar{X}_n|^2}{4\sigma_n^2}\right)\right] \qquad (A.14)$$

to obtain

$$E[a^2] = \frac{|\bar{X}_n|^2}{2}\left[1 + \frac{I_1\left[\frac{|\bar{X}_n|^2}{4\sigma_n^2}\right]}{I_0\left[\frac{|X_n|^2}{4\sigma_n^2}\right]}\right] + \sigma_n^2 \qquad (A.15)$$

Subtracting $E^2[a]$ gives the tabulated variance.

Integrating the expression for $p(a,\phi)$ over a gives

$$p(\phi) = \begin{cases} \dfrac{\exp\left\{\left[-|\bar{X}_n|^2/4\sigma_n^2\right]\left[1 - 2\cos^2(\phi - \delta_n)\right]\right\}}{2\pi\, I_0\left[\frac{|\bar{X}_n|^2}{4\sigma_n^2}\right]}\left[1 + \operatorname{erf}\dfrac{|X_n|\cos(\phi - \delta_n)}{\sigma_n}\right], & -\pi \le \phi \le \pi \\ 0, & \text{otherwise.} \end{cases} \qquad (A.16)$$

with erf(x) = the error function. No closed-form expressions exist for the moments of this density, so efforts are confined to finding large- and small-sample equations. First, however, the mean and variance must be computed for the final simple reproducing-type density, the density for learning W for a rectangular distribution (Case 10).

E[W] is found by straightforward integration

$$E[W] = \int_{M_n}^{\infty} (n-1)\left(\frac{M_n}{W}\right)^{n-1} dW, \qquad n > 1$$

$$= \begin{cases} \frac{n-1}{n-2} M_n, & n > 2, \\ \infty, & 1 < n \leq 2 \end{cases} \tag{A.17}$$

Similarly

$$E[W^2] = \int_{M_n}^{\infty} (n-1)\, M_n \left(\frac{M_n}{W}\right)^{n-2} dW, \qquad n > 1$$

$$= \begin{cases} \frac{n-1}{n-3} M_n, & n > 3, \\ \infty & 1 < n \leq 3 \end{cases} \tag{A.18}$$

Subtracting $E^2[W]$ gives Var [W] except for the case $1 < n \leq 2$, which is of the form $\infty - \infty$ and hence undefined.

C. LARGE-SAMPLE LIMITS OF MOMENTS

The limiting forms of many of the parameters in Table 4 may be obtained by the simple algebraic process of letting the size of the set of observations grow without bound, then computing the limits obtained. This process gives all of the values tabulated as zero in Table 4.

The limiting forms of most of the remainder of the parameters follow directly from application of the strong law of large numbers if the limits

are determined by actual observations. For the binomial distribution, Case 1:

$$E\left[\frac{r}{n}\middle|P = P_o\right] = P_o \tag{A.19}$$

Hence, by the strong law of large numbers

$$\frac{r}{n} \to P_o \tag{A.20}$$

with probability one.

Similar reasoning applies in most of the other cases studied. In case of the multinomial distribution, Case 2:

$$E\left[\frac{r_n}{n}\middle|P_i = P_{io}\right] = P_{io} \tag{A.21}$$

For the binary Markov Process, Case 3:

$$E\left[\frac{r_{ii}}{n_i}\middle|P_{ii} = P_{iio}\right] = P_{iio} \tag{A.22}$$

For the Poisson process, Case 4:

$$E\left[\frac{n}{\tau}\middle|\alpha = \alpha_o\right] = \alpha_o \tag{A.23}$$

For the Gaussian process with unknown mean vector, Case 5:

$$E[(\bar{\mathbf{X}}_n)_i | m_i = m_{io}] = m_{io} \tag{A.24}$$

or with unknown covariance matrix, Case 6:

$$E\left[\left(\frac{\mathbf{V}_n}{n-1}\right)^{ij} \middle| \mathbf{K}^{-1} = \mathbf{K}_o^{-1}\right] = k_o^{ij} \tag{A.25}$$

For the complex Gaussian process, Case 7:

$$E[|\bar{X}_n|\,\big|a = a_o] = a_o \tag{A.26}$$

and

$$E[\delta_n | \phi = \phi_o] = \phi_o \tag{A.27}$$

For the Rayleigh process, Case 8:

$$E\left[\frac{\Sigma X_i^2}{2n} \middle| \sigma^2 = \sigma_o^2\right] = \sigma_o^2 \tag{A.28}$$

hence, the reciprocal parameter ρ converges to $1/\sigma_o^2$. For the exponential process, Case 9:

$$E\left[\frac{\Sigma X_i}{n} \middle| \lambda = \lambda_o\right] = 1/\lambda_o \tag{A.29}$$

with the same type of reciprocal relationship as found in the corresponding case, Case 9, in Table 3.

In each of these cases, the strong law of large numbers applies in the same manner as in the binomial case. The only case differing is Case 10, the rectangular distribution. Convergence can be proved in this case also, but the proof differs from that for the other cases.

Since in Case 10 the sequence of M_n's is bounded and monotone, it must have a limit, with probability one. This limit must be W_o if the latter is the true value of W, since if the limit were not W_o it would have to be less than W_o. Then the Borel-Cantelli lemmas [Ref. 18] would state that values between the limit and W_o occurred infinitely often in an infinite sequence of observations, a contradiction. Hence, M_n must converge to W_o with probability one.

The limiting forms for means and variances in all cases save the complex Gaussian, Case 7, then follow immediately from Table 3. For the complex Gaussian density the limiting forms of the moments for a follow from expansion of the Bessel function terms, using the usual asymptotic expansions valid for large arguments [Ref. 28]. The moments of ϕ follow from the limiting form of $p(\phi)$:

$$p(\phi) \to \frac{|\bar{X}_n|}{\sqrt{2\pi}\,\sigma_n} \exp\left\{\left[-|\bar{X}_n|^2 \sin 2(\phi-\delta_n)\right] / 2\,\sigma_n^2\right\}, \quad -\pi \le \phi \le \pi \tag{A.30}$$

Since $\sigma_n^2 \to 0$, Expression (A.30) approaches zero except for $\phi \approx \delta_n$. Hence, $E[\phi] \to \delta_n$. The order of magnitude of the variance can be estimated from the width of the pulse given by Expression (A.30). This is obviously of the order of σ_n^2. The variance is a measure of the width of the pulse and must be of the same order of magnitude. The limiting form of the covariance is at most of the maximum order of the variances.

D. SMALL-SAMPLE LIMITS OF MOMENTS

The values of all limits in Table 5, save for the complex Gaussian case, phase variations, are obtained immediately from taking limits in Table 3. The moments for uniform densities may be found tabulated in the cases where the parameter range is finite. If the parameter range is infinite, and a function of θ is unbounded and non-negative, the limiting value of the expectation of the function, as the density on θ approaches uniformity, is infinite; while if the function can be both positive and negative, the limiting expectation is undefined. This gives all values in Table 5 save for the moments of ϕ in the seventh case.

For these moments of ϕ it is merely necessary to evaluate the expression for $p(\phi)$ in Eq. (A.16) as n approaches zero and σ_n^2 approaches infinity. The limit is a uniform density over the range $-\pi \leq \phi \leq \pi$.

E. LIKELIHOOD FOR COSINE OF UNKNOWN FREQUENCY

Section B of Chapter VII applied the learning technique to finding the expectation of a random variable--specifically a likelihood ratio involving a cosine of unknown frequency. It was necessary to integrate Eq. (83) twice to obtain Eq. (84). Since $p(\phi)$ is uniform over the range $[0,2\pi]$:

$$\ell(X|a,f) = \frac{\exp\left[-a^2 T_1/N_o\right]}{2\pi} \int_0^{2\pi} \exp\left\{\frac{4a}{N_o}\left[\int_0^{T_1} X(t)\cos(\omega t + \theta)\,dt\right]\right\} d\phi \qquad \text{(A.31)}$$

Expanding the cosine term,

$$\int_0^{T_1} X(t) \cos(\omega t+\phi)dt = \cos\phi \int_0^{T_1} X(t) \cos \omega t dt - \sin\phi \int_0^{T_1} X(t) \sin \omega t dt \tag{A.32}$$

Hence

$$\ell(X|a,f) = e^{-a^2 T_1/N_o} I_0\left[\frac{4a}{N_o}\left|\int_0^{T_1} X(t)\, e^{i\omega t}dt\right|\right] \tag{A.33}$$

Then, since by hypothesis, a is Rayleigh-distributed with parameter A^2:

$$\ell(X|f) = \int_0^{\infty} \frac{a}{A^2} e^{-a^2/2N_o B^2} I_0\left[\frac{a}{N_o B^2} \cdot 4\left|\int_0^{T_1} X(t)\, e^{i\omega t}dt\right| B^2\right] da$$

$$= \frac{N_o B^2}{A^2} \exp\left\{\frac{8B^2}{N_o}\left|\int_0^{T_1} X(t)\, e^{i\omega t}dt\right|^2\right\} \tag{A.34}$$

wherein use is made of the fact that the integral of the Rician density is unity in a manner analogous to Section B of this Appendix.

REFERENCES

1. C. E. Shannon, "A Mathematical Theory of Communication," Bell Sys. Tech J., 27, 1948, pp. 379-423, 623-656.

2. C. E. Shannon, "Communication in the Presence of Noise," Proc. IRE, 37, 1949, pp. 10-21.

3. N. Wiener, Cybernetics, The Technology Press, Cambridge, Mass. and John Wiley and Sons, New York, 1948.

4. N. Wiener, Extrapolation, Interpolation and Smoothing of Stationary Time Series, The Technology Press, Cambridge, Mass. and John Wiley and Sons, New York, 1949.

5. D. V. Lindley, "The Use of Prior Probability Distributions in Statistical Inference and Decisions," Proc. of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1, University of California Press, Berkeley, 1961, pp. 453-468.

6. L. J. Savage, "The Foundations of Statistics Reconsidered," Proc. of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1, University of California Press, Berkeley, 1961, pp. 575-586.

7. D. Braverman, "Machine Learning and Automatic Pattern Recognition," TR No. 2003-1, Contract Nonr 225(24), Stanford Electronics Laboratories, Stanford, Calif. 17 Feb 1961.

8. D. Braverman, "Learning Filters for Optimum Pattern Recognition," IRE Trans. (Information Theory), IT-8, Jul 1962, pp. 280-285.

9. N. Abramson and D. Braverman, "Learning to Recognize Patterns in a Random Environment," TR No. 2003-5 (SEL-62-071), Stanford Electronics Laboratories, Stanford, Calif. May 1962. Also in IRE Trans. (Information Theory), IT-8, Sep 1962, pp. 58-63.

10. D. G. Keehn, "Learning the Mean Vector and Covariance Matrix of Gaussian Signals in Pattern Recognition," TR No. 2003-6 (SEL-62-155), Stanford Electronics Laboratories, Stanford, Calif. Feb 1963.

11. R. Bellman, Adaptive Control Processes: A Guided Tour, Princeton University Press, Princeton, N.J. 1961.

12. J. E. Mosimann, "On the Compound Multinomial Distribution, the Multivariate Beta-Distribution, and Correlations among Proportions," Biometrika, 49, 1962, pp. 65-82.

13. G. Turin, "Communication through Noisy, Random-Multipath Channels," TR No. 116, MIT Lincoln Laboratory, Cambridge, Mass. May 1956.

REFERENCES (Continued)

14. T. Kailath, "Adaptive Matched Filters," Mathematical Optimization Techniques, edited by R. Bellman, University of California Press, Berkeley, 1963, pp. 109-140.

15. H. Raiffa and R. Schlaifer, Applied Statistical Decision Theory, Graduate School of Business Administration, Harvard University, Cambridge, Mass. 1961.

16. R. F. Daly, "Adaptive Binary Detectors," TR No. 2003-2, Contract Nonr 225(24), Stanford Electronics Laboratories, Stanford, Calif. 26 Jun 1961.

17. R. F. Daly, "The Adaptive Binary-Detection Problem on the Real Line," TR No. 2003-3 (SEL-62-030), Stanford Electronics Laboratories, Stanford, Calif. Feb 1962.

18. M. Loéve, Probability Theory, 2nd Ed., D. Van Nostrand Co., Inc., Princeton, N. J. 1960.

19. E. B. Dynkin, "Necessary and Sufficient Statistics for a Family of Probability Distributions," Selected Translations in Mathematical Statistics and Probability, 1, 1961, pp. 17-40.

20. D. Blackwell and M. A. Girshick, Theory of Games and Statistical Decisions, John Wiley and Sons, Inc., New York, 1954.

21. A. Rényi, "On a New Axiomatic Theory of Probability," Acta Mathematica Hungaricae, 6, 1955, pp. 285-335.

22. S. Kullback, Information Theory and Statistics, John Wiley and Sons, Inc., New York, 1959.

23. T. L. Grettenberg, "A Criterion for the Statistical Comparison of Communication Systems with Applications to Optimum Signal Selection," TR No. 2004-4 (SEL-62-013), Stanford Electronics Laboratories, Stanford, Calif. Feb 1962.

24. B. L. Basore, "A Model for Communication with Learning," TN-2-1004, Contract AF30(602)-1890, The Dikewood Corp., Albuquerque, N.M. 31 May 1960. ASTIA AD No. 242535.

25. B. L. Basore, "Communication Applications of Information Theory," DFR-1004, Contract AF30(602)-1890, The Dikewood Corp., Albuquerque, N.M. 29 Dec 1960. ASTIA AD No. 257983.

26. S. Watanabe, "Information Theoretical Aspects of Inductive and Deductive Inference," IBM J. of Res. and Dev., 4, 2, Apr 1960.

27. D. T. Magill, "Optimal Adaptive Estimation of Sampled Stochastic Processes," to be published as a Stanford Electronics Laboratories technical report.

28. W. G. Bickley, Bessel Functions and Formulae, Cambridge University Press, Cambridge, England, 1957.

29. S. S. Wilks, Mathematical Statistics, John Wiley and Sons, New York, 1962.

30. E. Parzen, Modern Probability Theory and Its Applications, John Wiley and Sons, New York, 1960.

31. T. W. Anderson, An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, New York, 1958.

32. D. Middleton, An Introduction to Statistical Communication Theory, McGraw-Hill Book Co., Inc., New York, 1960.

SYSTEMS THEORY
DISTRIBUTION LIST
September 1963

GOVERNMENT

USAELRDL
Ft. Monmouth, N.J.
1 Attn: Dr. H. Jacobs, SIGRA/SL-PF

Procurement Data Division
USAF Equipment Support Agency
Ft. Monmouth, N.J.
1 Attn: Mr. M. Rosenfeld

Commanding General, USAELRDL
Ft. Monmouth, N.J.
SIGRA SL-SC, Bldg. [illegible]
1 TDC, Evans Signal Lab Area

Commanding Officer, ERDL
Ft. Belvoir, Va.
1 Attn: Tech. Doc. Ctr.

Commanding Officer
Frankford Arsenal
Bridge and Tacony St.
Philadelphia [illegible], Pa.
1 Attn: Library Br., [illegible], Bldg. 40

Ballistics Research Lab
Aberdeen Proving Ground, Md.
Attn: V. W. [illegible], BML

Chief of Naval Research
Navy Dept.
Washington [illegible], D.C.
Attn: Code [illegible]
1 Code 420
1 Code [illegible]

Commanding Officer, USAEPU
P.O. Box [illegible]
1 Mt. View, Calif.

Commanding Officer
ONR Branch Office
[illegible] Geary St.
1 San Francisco [illegible], Calif.

Commanding Officer
ONR Branch Office
[illegible] E. Green St.
1 Pasadena, Calif.

Office of Naval Research
Branch Office Chicago
[illegible] N. Michigan Ave.
1 Chicago 1, Ill.

Commanding Officer
ONR Branch Office
[illegible]
1 New York 11, N.Y.

Office of Naval Research
Navy 100, Box [illegible]
Fleet Post Office
New York, N.Y.
1 Attn: Dr. I. Rowe

New York Naval Shipyard
Material Laboratory Library
Brooklyn, N.Y.
1 Attn: Code 911B, M. Kosofsky
Bldg. [illegible]

Chief Bureau of Ships
Navy Dept.
Washington [illegible], D.C.
1 Attn: Code [illegible]A1
1 Code [illegible]
1 Code [illegible] NIDS
1 Code [illegible] D
1 Code [illegible], A. E. Smith
1 Code [illegible] 1A

Officer in Charge, ONR
Navy 100, Box 39, Fleet P.O.
10 New York, N.Y.

U.S. Naval Research Lab
Washington 25, D.C.
6 Attn: Code 2000
1 5240
1 5430
1 5200
1 5300
1 5400
1 5266, G. Abraham
1 2027
1 5260
1 6430

Chief, Bureau of Naval Weapons
Navy Dept.
Washington 25, D.C.
1 Attn: RAAV-6
1 RUDC-1
2 RREN-3
1 RAAV-44

Chief of Naval Operations
Navy Dept.
Washington 25, D.C.
1 Attn: Code Op 945Y

Director, Naval Electronics Lab
1 San Diego 52, Calif.

USN Post Graduate School
1 Monterey, Calif.
1 Attn: Tech. Reports Librarian
1 Prof. Gray, Electronics Dept.
1 Dr. H. Titus

Weapons Systems Test Div.
Naval Air Test Center
Patuxent River, Md.
1 Attn: Library

U.S. Naval Weapons Lab
Dahlgren, Va.
1 Attn: Tech. Library

Naval Ordnance Lab
Corona, Calif.
1 Attn: Library
1 H. H. Wieder, 423

Commander, USN Air Dev. Ctr.
Johnsville, Pa.
1 Attn: NADC Library
1 AD-5

Commander
USN Missile Center
Pt. Mugu, Calif.
1 Attn: N03022

Commanding Officer
U.S. Army Research Office
Box CM, Duke Station
Durham, N.C.
3 Attn: CRD-AA-IP

Commanding General
U.S. Army Materiel Command
Washington 25, D.C.
1 Attn: AMCRD-DE-E
1 AMCRD-RS-PE-E

Department of the Army
Office, Chief of Res. and Dev.
The Pentagon
Washington 25, D.C.
1 Attn: Research Support Div., Rm. 3D442

Office of the Chief of Engineers
Dept. of the Army
Washington 25, D.C.
1 Attn: Chief, Library Br.

Office of the Asst. Secy. of Defense
Washington 25, D.C.
1 (AE) Pentagon Bldg., Rm. 3D984

Hqs., USAF(AFRDR-NU.3)
The Pentagon, Washington 25, D.C.
1 Attn: Mr. H. Mulkey, Rm. 4D335

Chief of Staff, USAF
Washington 25, D.C.
2 Attn: AFDRT-ER

Hq., USAF
Dir. of Science and Technology
Electronics Div.
Washington 25, D.C.
1 Attn: AFRST-EL/CS, Maj. E. N. Myers

Aeronautical Systems Div.
Wright-Patterson AFB, Ohio
1 Attn: Lt. Col. L. M. Butsch, Jr.
ASRNE-2
1 ASRNE-2, D. R. Moore
1 ASRNR-32
1 ASRNE-1, Electronic Res. Br. Elec. Tech. Lab
1 ASRNCF-2, Electromagnetic and Comm. Lab
3 ASNXR
1 ASNXRR(Library)
1 ASRNE-32

Commandant
AF Institute of Technology
Wright-Patterson AFB, Ohio
1 Attn: AFIT (Library)

Executive Director
AF Office of Scientific Res.
Washington 25, D.C.
1 Attn: SRYA

AF Special Weapons Center
Kirtland AFB, N.M.
2 Attn: SWOI

Director
Air University Library
Maxwell AFB, Ala.
1 Attn: CR-4582

AF Missile Test Center
Patrick AFB, Fla.
1 Attn: AFMTC Tech. Library, MU-135

Commander, AF Cambridge Res. Labs
ARDC, L. G. Hanscom Field
Bedford, Mass.
1 Attn: CRTOTT-2, Electronics

Hqs., AF Systems Command
Andrews AFB
Washington 25, D.C.
1 Attn: SCTAE

Asst. Secy. of Defense (R and D)
R and D Board, Dept. of Defense
Washington 25, D.C.
1 Attn: Tech. Library

Office of Director of Defense
Dept. of Defense
Washington 25, D. C.
1 Attn. Research and Engineering

Institute for Defense Analyses
1666 Connecticut Ave.
Washington 9, D. C.
1 Attn: W. E. Bradley

Defense Communications Agency
Dept. of Defense
Washington 25, D. C.
1 Attn: Code 101A, Tech. Library

Advisory Group on Electron Devices
346 Broadway, 8th Floor East
New York 13, N. Y.
1 Attn: H. Sullivan

Advisory Group on Reliability of
Electronic Equipment
Office Asst. Secy. of Defense
The Pentagon
1 Washington 25, D.C.

Commanding Officer
Diamond Ordnance Fuze Labs
Washington 25, D. C.
1 Attn: ORDTL 930, Dr. R. T. Young

Diamond Ordnance Fuze Lab.
U.S. Ordnance Corps
Washington 25, D. C.
1 Attn: ORDTL-430-638,
Mr. R. H. Comyn

U.S. Dept. of Commerce
National Bureau of Standards
Boulder Labs
Central Radio Propagation Lab.
1 Boulder, Colorado
1 Attn: Miss J. V. Lincoln, Chief
RWPG

NAF, Engineering Section
1 Washington, D.C.

Information Retrieval Section
Federal Aviation Agency
Washington, D. C.
1 Attn: MS-11, Library Branch

DDC
Cameron Station
Alexandria 4, Va.
10 Attn: TISIA

U.S. Coast Guard
1300 E. Street, N.W.
Washington 25, D. C.
1 Attn: EEE Station 540

Office of Technical Services
Dept. of Commerce
1 Washington 25, D.C.

Director
National Security Agency
Fort George G. Meade, Md.
1 Attn: R4

NASA, Goddard Space Flight Center
Greenbelt, Md.
1 Attn: Code 611, Dr. G. H. Ludwig
1 Chief, Data Systems Division

Chief, U.S. Army Security Agency
Arlington Hall Station
1 Arlington 12, Virginia

*No AF or Classified Reports.

SCHOOLS

*U of Aberdeen
Dept. of Natural Philosophy
Marischal College
Aberdeen, Scotland
1 Attn: Mr. R. V. Jones

U of Arizona
EE Dept.
Tucson, Ariz.
1 Attn: R. L. Walker
1 D. J. Hamilton

*U of British Columbia
Vancouver 8, Canada
1 Attn: Dr. A. C. Soudack

California Institute of Technology
Pasadena, Calif.
1 Attn: Prof. R. W. Gould
1 Prof. L. M. Field, EE Dept.
1 D. Braverman, EE Dept.

California Institute of Technology
4800 Oak Grove Drive
Pasadena 3, Calif.
1 Attn: Library, Jet Propulsion Lab.

U. of California
Berkeley 4, Calif.
1 Attn: Prof. R. M. Saunders, EE Dept.
1 Dr. R. K. Wakerling,
Radiation Lab. Info. Div.,
Bldg. 30, Rm. 101

U of California
Los Angeles 24, Calif.
1 Attn: C. T. Leondes, Prof. of
Engineering, Engineering
Department
1 R. S. Elliott,
Electromagnetics Div., College
of Engineering

U of California, San Diego
School of Science and Engineering
La Jolla, Calif.
1 Attn: Physics Dept.

Carnegie Institute of Technology
Schenley Park
Pittsburg 13, Pa.
1 Attn: Dr. E. M. Williams, EE Dept.

Case Institute of Technology
Engineering Design Center
Cleveland 6, Ohio
1 Attn: Dr. J. B. Neswick, Director

Cornell U
Cognitive Systems Research Program
Ithaca, N. Y.
1 Attn: F. Rosenblatt, Hollister Hall

Drexel Institute of Technology
Philadelphia 4, Pa.
1 Attn: F. B. Haynes, EE Dept.

U of Florida
Engineering Bldg., Rm. 330
Gainsville, Fla.
1 Attn: M. J. Wiggins, EE Dept.

Georgia Institute of Technology
Atlanta 13, Ga.
1 Attn: Mrs. J. H. Crosland
Librarian
1 F. Dixon, Engr. Experiment
Station

Harvard U
Pierce Hall
Cambridge 38, Mass.
1 Attn: Dean H. Brooks, Div of Engr.
and Applied Physics, Rm. 217
2 E. Parker, Librarian, Rm.
303A, Tech. Reports
Collection

U of Hawaii
Honolulu 14, Hawaii
1 Attn: Asst. Prof. K. Najita,
EE Dept.

Illinois Institute of Technology
Technology Center
Chicago 16, Ill.
1 Attn: Dr. P. C. Yuen, EE Dept.

U of Illinois
Urbana, Ill.
1 Attn: P. D. Coleman, EE Res. Lab.
1 W. Perkins, EE Res. Lab.
1 A. Albert, Tech. Ed., EE
Res. Lab.
1 Library Serials Dept.
1 Prof. D. Alpert, Coordinated
Sci. Lab.

*Instituto de Pesquisas da Marinha
Ministerio da Marinh
Rio de Janeiro
Estado da Guanabara, Brazil
1 Attn: Roberto B. da Costa

Johns Hopkins U
Charles and 34th St.
Baltimore 18, Md.
1 Attn: Librarian, Carlyle Barton Lab.

Johns Hopkins U
6621 Georgia Ave.
Silver Spring, Md.
1 Attn: H. H. Choksy
1 Mr. A. W. Love, Applied
Physics Lab.

Linfield Research Institute
McMinnville, Ore.
1 Attn: G. M. Hekok, Director

Marquette University
College of Engineering
1131 W. Wisconsin Ave.
Milwaukee 3, Wis.
1 Attn: A. C. Mueller, EE Dept.

M I T
Cambridge 39, Mass.
1 Attn: Res. Lab. of Elec., Doc.
Rm. 36-412
1 Miss A. Ellis, Libn., Rm 4-144,
LIB
1 Mr. J. E. Ward, Elec. Sys.
Lab.

M I T
Lincoln Laboratory
P.O. Box 73
1 Attn: Lexington 73, Mass.
1 Navy Representative
1 Dr. W. I. Wells

U of Michigan
Ann Arbor, Mich.
1 Attn: Dir., Cooley Elec. Labs.,
N. Campus
1 Dr. J. E. Rowe, Elec. Engr.
Lab.
1 Comm. Sci. Lab., 1200 Frieze
Bldg.

U of Michigan
Institute of Science and Technology
P.O. Box 618
Ann Arbor, Mich.
1 Attn: Tech. Documents Service
1 W. Wolfe--IRIA--

U of Minnesota
Institute of Technology
Minneapolis 14, Minn.
1 Attn: Prof. A. Van der Ziel,
EE Dept.

U. of Nevada
College of Engineering
Reno, Nev.
1 Attn: Dr. R. A. Marbart, EE Dept.

Northeastern U
The Dodge Library
Boston 15, Mass.
1 Attn: Jerome B. Locke, Librarian

Northwestern U
[illegible] Oakton St.
Evanston, Ill.
1 Attn: W. C. Toth, Aerial
Measurements Lab.

U of Notre Dame
South Bend, Ind.
1 Attn: [illegible], EE Dept.

Ohio State U
[illegible] Kinnear Ave.
Columbus 10, Ohio
1 Attn: Prof. L. M. Boone, EE Dept.

Oregon State U
Corvallis, Ore.
1 Attn: H. J. Oberthaler, EE Dept.

Polytechnic Institute
[illegible] Jay St.
Brooklyn, N. Y.
1 Attn: L. Shaw, EE Dept.

Polytechnic Institute of Brooklyn
Graduate Center, Route 110
Farmingdale, N. Y.
1 Attn: Librarian

Purdue U
Lafayette, Ind.
1 Attn: Librar., EE Dept.

Rensselaer Polytechnic Institute
Troy, N. Y.
1 Attn: Librar., Serials Dept.

*U of Saskatchewan
College of Engineering
Saskatoon, Canada
1 Attn: Prof. R. E. Ludwig

Stanford Research Institute
Menlo Park, Calif.
1 Attn: Material Reports, G-037

Syracuse U
Syracuse 10, N. Y.
1 Attn: EE Dept.

*Uppsala U
Institute of Physics
Uppsala, Sweden
1 Attn: Dr. P. A. Tove

U of Utah
Salt Lake City, Utah
1 Attn: R. W. Grow, EE Dept.

U of Virginia
Charlottesville, Va.
1 Attn: J. C. Wyllie, Alderman
Library

U of Washington
Seattle 5, Wash.
1 Attn: A. E. Harrison, EE Dept.

Worchester Polytechnic Inst.
Worchester, Mass.
1 Attn: Dr. H. H. Newell

Yale U
New Haven, Conn.
1 Attn: Sloane Physics Lab.
1 EE Dept.
1 Dunham Lab., Engr. Library

INDUSTRIES

Argonne National Lab.
9700 South Cass
Argonne, Ill.
1 Attn: Dr. O. C. Simpson

Admiral Corp.
3800 Cortland St.
Chicago 47, Ill.
1 Attn: E. N. Roberson, Librarian

Airborne Instruments Lab.
Comac Road
Deer Park, Long Island, N. Y.
1 Attn: J. Dyer, Vice-Pres. and
Tech. Dir.

Amperex Corp.
230 Duffy Ave.
Hicksville, Long Island, N. Y.
1 Attn: Proj. Engineer, S. Barbasso

Autonetics
Div. of North American Aviation, Inc
9150 E. Imperial Highway
Downey, Calif.
1 Attn: Tech. Library 3040-3

Bell Telephone Labs.
Murray Hill Lab.
Murray Hill, N. J.
1 Attn: Dr. J. R. Pierce
1 Dr. S. Darlington
1 Mr. A. J. Grossman

Bell Telephone Labs., Inc.
Technical Information Library
Whippany, N. J.
1 Attn: Tech. Repts. Libr.,
Whippany Lab.

*Central Electronics Engineering
Research Institute
Pilani, Rajasthan, India
1 Attn: Om P. Gandhi - Via:
ONR/London

Columbia Radiation Lab.
538 West 120th St.
1 New York, New York

Convair - San Diego
Div. of General Dynamics Corp.
San Diego 12, Calif.
1 Attn: Engineering Library

Cook Research Labs.
6401 W. Oakton St.
1 Attn: Morton Grove, Ill.

Cornell Aeronautical Labs., Inc.
4455 Genesee St.
Buffalo 21, N. Y.
1 Attn: Library

Eitel-McCullough, Inc.
301 Industrial Way
San Carlos, Calif.
1 Attn: Research Librarian

Ewan Knight Corp.
East Natick, Mass.
1 Attn: Library

Fairchild Semiconductor Corp.
4001 Junipero Serra Blvd.
Palo Alto, Calif.
1 Attn: Dr. V. H. Grinich

General Electric Co.
Defense Electronics Div., LMED
Cornell University, Ithaca, N. Y.
1 Attn: Library - Via: Commander,
AGD W-P AFB, Ohio, ASRNGW
D.E. Lewis

General Electric TWT Products Sec.
601 California Ave.
Palo Alto, Calif.
1 Attn: Tech. Library, C. G. Lob

General Electric Co. Res. Lab
P.O. Box 1088
Schenectady, N.Y.
1 Attn: Dr. P. M. Lewis
1 R. L. Shuey, Mgr. Info.
Studies Sec.

General Electric Co.
Electronics Park
Bldg. 3, Rm. 143-1
Syracuse, N.Y.
1 Attn: Doc. Library, Y. Burke

Gilfillan Brothers
1815 Venice Blvd.
Los Angeles, Calif.
1 Attn: Engr. Library

The Hallicrafters Co.
5th and Kostner Ave.
1 Attn: Chicago 24, Ill.

Hewlett-Packard Co.
1501 Page Mill Road
1 Attn: Palo Alto, Calif.

Hughes Aircraft
Malibu Beach, Calif.
1 Attn: Mr. Iams

Hughes Aircraft Co.
Florence at Teale St.
Culver City, Calif.
1 Attn: Tech. Doc. Cen., Bldg. 6,
Rm. C1043

Hughes Aircraft Co.
P.O. Box 278
Newport Beach, Calif.
1 Attn: Library, Semiconductor Div.

IBM, Box 390, Boardman Road
Poughkeepsie, N. Y.
1 Attn: J. C. Logue, Data Systems Div.

IBM, Poughkeepsie, N. Y.
1 Attn: Product Dev. Lab., E. M.
Davis

*No Air or Classified reports.

IBM ASD and Research Library
Monterey and Cottle Roads
San Jose, Calif.
1 Attn: Miss M. Griffin, Bldg. 025

ITT Federal Labs.
500 Washington Ave.
Nutley 10, N. J.
1 Attn: Mr. E. Mount, Librarian

Laboratory for Electronics, Inc.
1075 Commonwealth Ave.
Boston 15, Mass.
1 Attn: Library

LEL, Inc.
75 Akron St.
Copiague, Long Island, N. Y.
1 Attn: Mr. R. S. Mariner

Lenkurt Electric Co.
San Carlos, Calif.
1 Attn: M. L. Waller, Librarian

Librascope
Div. of General Precision, Inc.
808 Western Ave.
Glendale 1, Calif.
1 Attn: Engr. Library

Lockheed Missiles and Space Div.
P.O. Box 504, Bldg. 52
Sunnyvale, Calif.
1 Attn: Dr. W. M Harris, Dept. 57-30
1 G. W. Price, Dept. 57-33

Melpar, Inc.
3000 Arlington Blvd.
Falls Church, Va.
1 Attn: Librarian

Microwave Associates, Inc.
Northwest Industrial Park
Burlington, Mass.
1 Attn: K. Mortenson
1 Librarian

Microwave Electronics Corp.
4061 Transport St.
Palo Alto, Calif.
1 Attn: S. F. Kaisel
1 M. C. Long

Minneapolis-Honeywell Regulator Co.
1177 Blue Heron Blvd.
Riviera Beach, Fla.
1 Attn: Semiconductor Products Library

Monsanto Research Corp.
Station B, Box 8
Dayton 7, Ohio
1 Attn: Mrs. D Crabtree

Monsanto Chemical Co.
800 N. Linbergh Blvd.
St. Louis 66, Mo
1 Attn: Mr. E. Orban, Mgr. Inorganic Dev.

*Dir., National Physical Lab.
Hillside Road
New Delhi 12, India
1 Attn: S.C. Sharma - Via: ONR/London

*Norther Electric Co., Lmtd.
Research and Development Labs.
P.O. Box 3511, Station "C"
Ottawa, Ontario, Canada
1 Attn: J. F. Tatlock
Via: ASD, Foreign Release Office
W-P AFB, Ohio
Mr. J. Troyan (ASYF)

Nortronics
Palo Verdes Research Park
6101 Crest Road
Palos Verdes Estates, Calif.
1 Attn: Tech. Info. Center

Pacific Semiconductors, Inc.
14520 So. Aviation Blvd.
Lawndale, Calif.
1 Attn: H. G. North

Philco Corp.
Tech. Rep. Division
P.O. Box 4730
Philadelphia 34, Pa.
1 Attn: F. R. Sherman, Mgr. Editor

Philco Corp.
Jolly and Union Meeting Roads
Blue Bell, Pa.
1 Attn: C. T. McCoy
1 Dr. J. R. Feldmeier

Polarad Electronics Corp.
43-20 Thirty-Fourth St.
Long Island City 1, N. Y.
1 Attn: A. H. Sonnenschein

Radio Corp. of America
RCA Labs., David Sarnoff Res. Cen.
Princeton, N. J.
2 Attn: Dr. J. Sklansky

RCA Labs., Princeton, N. J.
1 Attn: H. Johnson

RCA, Missile Elec. and Controls Dept.
Woburn, Mass.
1 Attn: Library

The Rand Corp.
1700 Main St.
Santa Monica, Calif.
1 Attn: Helen J. Waldron, Librarian

Raytheon Manufacturing Co.
Microwave and Power Tube Div.
Burlington, Mass.
1 Attn: Librarian, Spencer Lab.

Raytheon Manufacturing Co.
Res. Div., 28 Seyon St.
Waltham, Mass.
1 Attn: Dr. H. Statz
1 Mrs. M. Bennett, Librarian

Roger White Electron Devices, Inc.
Tall Oaks Road
1 Laurel Hedges, Stamford, Conn.

Sandia Corp.
Sandia Base, Albuquerque, N. M.
1 Attn: Mrs. B. R. Allen, Librarian

Sperry Rand Corp.
Sperry Electron Tube Div.
1 Gainesville, Fla.

Sperry Gyroscope Co.
Div. of Sperry Rand Corp.
Great Neck, N.Y.
1 Attn: L. Swern(MS3T105)

Sperry Gyroscope Co.
Engineering Library
Mail Station F-7
Great Neck, Long Island, N. Y.
1 Attn: K. Barney, Engr. Dept. Head

Sperry Microwave Electronics
Clearwater, Fla.
1 Attn: J. E. Pippin, Res. Sec. Head

Sylvania Electric Products, Inc.
208-20 Willets Point Blvd.
Bayside, Long Island, N. Y.
1 Attn: L. R. Bloom, Physics Lab.

Sylvania Electronics Systems
100 First Ave.
Waltham 54, Mass.
1 Attn: Librarian, Waltham Labs.
1 Mr. E. E. Hollis

Technical Research Group
1 Syosett, L.I., N.Y.

Texas Instruments, Inc.
Semiconductor-Components Div.
P.O. Box 205
Dallas 22, Tex.
1 Attn: Library
2 Dr. W. Adcock

Texas Instruments, Inc.
1300 N. Central Expressway
1 Dallas, Texas

Texas Instruments, Inc.
P.O. Box 6015
Dallas 22, Tex.
1 Attn: M. E. Coun, Apparatus Div.

Texas Instruments
6017 E. Calle Tuberia
Phoenix, Arizona
1 Attn: R. L. Pritchard

Texas Instruments, Inc.
Corporate Research and Engineering
Technical Reports Service
P.O. Box 5474
1 Attn: Dallas 22, Tex.

Textronix, Inc.
P.O. Box 500
Beaverton, Ore.
4 Attn: Dr. C. F. DeLord, Dir. of Research

Varian Associates
611 Hansen Way
Palo Alto, Calif.
1 Attn: Tech. Library

Weiterman Electronics
4649 North 38th St.
1 Milwaukee 9, Wisconsin

Westinghouse Electric Corp.
Friendship International Airport
Box 746, Baltimore 3, Md.
1 Attn: G. R. Kilgore, Mgr. Appl. Res. Dept. Baltimore Lab.

Westinghouse Electric Corp.
3 Gateway Center
Pittsburgh 22, Pa.
1 Attn: Dr. G. C. Sziklai

Westinghouse Electric Corp.
P.O. Box 284
Elmira, N. Y.
1 Attn: S. S. King

Zenith Radio Corp.
6001 Dickens Ave.
Chicago 39, Ill.
1 Attn: J. Markin

*No AF or Classified Reports.